How different visualizations affect human reasoning about uncertainty: An analysis of visual behaviour

Manuele Reani, Niels Peek, Caroline Jay

PII: S0747-5632(18)30523-5
DOI: https://doi.org/10.1016/j.chb.2018.10.033
Reference: CHB 5767

To appear in: Computers in Human Behavior

Received Date: 7 June 2018
Revised Date: 4 October 2018
Accepted Date: 23 October 2018

Please cite this article as: Reani M., Peek N. & Jay C., How different visualizations affect human reasoning about uncertainty: An analysis of visual behaviour, Computers in Human Behavior (2018), doi: https://doi.org/10.1016/j.chb.2018.10.033.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.
How Different Visualizations Affect Human Reasoning About Uncertainty: An Analysis of Visual Behaviour

Authors: Manuele Reani (a)*, Niels Peek (b,c), Caroline Jay (a)

(a) School of Computer Science, the University of Manchester, Kilburn Building, Oxford Road, Manchester M13 9PL, UK
(b) Health eResearch Centre, Division of Informatics, Imaging & Data Sciences, The University of Manchester
(c) NIHR Greater Manchester Patient Safety Translational Research Centre, School of Health Sciences, The University of Manchester
* Corresponding author: m.reani@manchester.ac.uk

Acknowledgements
MR’s work was funded by the Engineering and physical Science Research Council (EPSRC number: 1703971). NP’s work was partially funded by the Engineering and Physical Sciences Research Council (EP/P010148/1) and by the National Institute for Health Research (NIHR) Greater Manchester Patient Safety Translational Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Declarations of interest: none
How Different Visualizations Affect Human Reasoning About
Uncertainty: An Analysis of Visual Behaviour

ABSTRACT
Humans find reasoning about uncertainty difficult. In decision support systems and
software for intelligence analysis, graphical representations are commonly used to
display uncertainty. Nevertheless, our understanding of how people use the infor-
mation presented in graphs displaying uncertainty to make decisions is limited. As
many artificial intelligent systems require a human-in-the-loop who is able to actively
take part in the analysis process, the understanding of high-level cognition involved
in human-graph interaction is essential in the design of better tools for analysis.
In this research, we investigate the visual behaviour that is associated with par-
ticipants responses to problems testing probabilistic reasoning represented through
two different visualizations (tree and Venn diagrams). Using the data from visual
fixations and transitions, we present a description of different reasoning strategies
covering both accurate and inaccurate reasoning for different visualization formats.
The results show that gaze behaviour is related to reasoning accuracy. Moreover,
this study shows that different graphs representing the same problem evoke differ-
ton reasoning strategies, suggesting that higher level cognition is influenced by the
graphical representation in which uncertainty is encoded.

KEYWORDS
Probabilistic reasoning; eye tracking; information visualization; decision making;
visual analytics; human-computer interaction

1. Introduction

1.1. Research Rationale
The advent of modern computing has been transforming the way in which people
make use of data. Nevertheless, human understanding and input is still important
for certain tasks, especially in situations in which the information is incomplete, or
where it is difficult to model the reality using mathematics (Holzinger 2016). Recently,
the concept of the ‘human-in-the-loop’ has been proposed as a new approach to in-
telligent systems, in which the actualization of complex computations is performed
by the machine, but the higher level decision-making is based on human judgment
(Holzinger 2016; Holzinger et al. 2016, 2017; Doran, Schulz, and Besold 2017). This
stream of research has proposed new guidelines for the development of interactive tools
for data analysis that shift focus from a ‘black-box’ to a ‘transparent box’ viewpoint
of the analysis process. This approach can improve the outcomes of complex analysis
by increasing human-computer interactivity and it increases the transparency of the
analysis processes. In a society where regulations around data are becoming tighter,
having a human-in-the-loop who is able to understand and explain the results of com-
plex analysis to others is vital (Holzinger 2016; Holzinger et al. 2016, 2017; Doran,
Schulz, and Besold 2017). The number of decision support systems has vastly in-
creased in the past decade. These systems often rely on visualizations to provide users
with a quick summary of complex information (Görg et al. 2013; Kang and Stasko
2014; Sauter et al. 2011; Zhu and Chen 2008; Cota, Castro, and Domínguez 2014; Yu, Deng, and Zhang 2009). In particular, visualizations can be used to quantify risk in a certain situation, and aim to provide insight into uncertainty within the data. Understanding how human reason about uncertainty is therefore fundamental to designing useful visualizations. The aim of this research is to understand the cognitive processes underpinning probabilistic reasoning in the context of human-graph interaction, via an eye tracking method that uses a novel form of fixation and transition analysis.

1.2. Probabilistic Reasoning

Faulty reasoning about uncertainty can lead to poor judgment and suboptimal decision-making (Kahneman and Tversky 1973). Previous research shows that people tend to perform poorly in tasks requiring an understanding of probability. For instance, humans find Bayesian reasoning problems, which generally ask for a probabilistic estimate given existing knowledge about a situation, particularly hard to solve (Eddy 1982). An example of this, adapted from Brase (2014), asks people to predict the weather using low barometric pressure as an indicator of rain as follows:

In Gotham city, it is either rainy or sunny. On average, 10% of the days are rainy days. 90% of rainy days are associated with low barometric pressure. 20% of sunny days are also associated with low barometric pressure. What is the probability that today will be rainy, given that it is a low pressure day?

Arriving at the correct answer – which is 0.33 (33%) – requires the use of Bayes’ theorem, shown in equation 1, where P(H | E) is the probability of the hypothesis (it is a rainy day) given the evidence (it is a low pressure day), P(E | H) is the probability of the evidence (it is a low pressure day) given the hypothesis (it is a rainy day), P(H) is the probability of the hypothesis (it is a rainy day), P(E | H¬) is the probability of the evidence (it is a low pressure day) given the opposite hypothesis (it is a sunny day) and P(H¬) is the probability of the opposite hypothesis (it is a sunny day).

\[
P(H | E) = \frac{P(E | H) P(H)}{P(E | H) P(H) + P(E | H¬) P(H¬)} \tag{1}
\]


There is evidence that people may fail to consider the base rate information i.e. P(H) in the example above—a phenomenon known as base rate neglect (Bar-Hillel 1980); and that participants performance is affected by the magnitude of the base rate information provided in the problem (Reani et al. 2018). There is also evidence of an association between Bayesian reasoning ability and individual differences in numeracy and spatial ability (Brase and Hill 2017; Garcia-Retamero, Cokely, and Hoffrage 2015; Ottley et al. 2016). Theories of how people perform Bayesian reasoning include the
ecological rationality framework, stating that humans have evolved to reason about uncertainty by sampling natural frequencies in the real world, and the nested-set hypothesis stating that humans are better at reasoning when problems are presented in formats that clearly expose the relationships between sets (McDowell and Jacobs 2017). At present, however, details of the cognitive mechanisms underpinning the full Bayesian reasoning process remain elusive.

One of the most consistent findings in Bayesian reasoning research is that by reframing the problem using natural frequencies instead of percentages (or probabilities) one obtains a considerably larger number of correct inferences (McDowell and Jacobs 2017). A version of the weather problem presented above reframed using the natural frequency approach proposed by Gigerenzer and Hoffrage (1995) is as follows:

*In Gotham city, it is either rainy or sunny. On average, out of 1000 days, 100 are rainy days. Out of these 100 rainy days, 90 are associated with low barometric pressure. Out of the 900 remaining sunny days, 180 are also associated with low barometric pressure. Here is a new representative sample of days in Gotham city that are associated with low barometric pressure. How many of these days would you expect to be rainy?*

To solve this version of the problem one needs simply to divide the number of rainy days associated with low barometric pressure (Rainy-low = 90) by the total number days that are associated with low barometric pressure (Low = 90 + 180). Nevertheless, around 50% of people still make an incorrect inference, even when the problem is framed using this format (Gigerenzer and Hoffrage 1995).

Previous research shows that in the probability version of Bayesian reasoning problems, people use a number of incorrect strategies associated with poor reasoning, and a number of suboptimal strategies that, although incorrect, can lead to an approximately correct response in certain situations (Gigerenzer and Hoffrage 1995; Cohen and Staub 2015). For instance, a suboptimal strategy that is found only in the probability format is the rare event shortcut. This correspond to the omission of P(H¬) (e.g., the probability of being sunny) in the denominator of equation 1. This strategy results in a close-to-correct answer only if the base rate P(H) (i.e., the probability of rain) is very small (e.g., 0.001) which implies that P(H¬) is a number close to one (e.g., 0.999). In this case, multiplying P(E | H¬) by P(H¬) would not change P(E | H¬) much. This strategy is not used in the frequency format because the conjunction P(E & H¬), which is provided by the problem when using this format, already takes into account this multiplication (Gigerenzer and Hoffrage 1995). Although, the frequency format reduces the number of incorrect responses, people may still misinterpret the information. An example of a poor reasoning strategy that occurs in both the frequency and the probability formats is the comparison shortcut, in which one divides P(E & H) (i.e., the number of rainy days with low barometric pressure) by P(E & H¬) (i.e., the number of sunny days with low barometric pressure) – in the weather problem shown above, this is 90/180 (Gigerenzer and Hoffrage 1995). This strategy can lead to a close-to-correct answer only in cases where P(E & H¬) is much larger than P(E & H). However, if this is not the case, this strategy produces an answer very far from the true Bayesian inference, as demonstrated in the weather problem where 90/270 ≠ 90/180. There is some consistency across participants in the observed usage of incorrect strategies that do not approximate to a true response, regardless of the data provided in the problem, with both the probability and the frequency formats (Gigerenzer and Hoffrage 1995; Cohen and Staub 2015). For instance, the multiply-all
strategy consists of multiply \( P(E \& H) \) by \( P(E) \). This always produces an incorrect result and has been observed in both the probability and frequency formats (Gigerenzer and Hoffrage 1995).

How data is presented plays a key role in supporting reasoning about uncertainty (Spiegelhalter, Pearson, and Short 2011; Brase 2014). The use of visualization in Bayesian reasoning research has been extensively studied as it appears to facilitate the reasoning process (Brase 2014). Methods tend to emphasize frequencies, or nested set-relationships, or both (Brase 2014; Ottley et al. 2016; Khan et al. 2015; Reani et al. 2018). However, it remains unclear which is the optimal format for supporting reasoning, not least because we still do not fully understand the cognitive processes these formats support (McDowell and Jacobs 2017).

1.3. Research Objective and Hypotheses

In this research we use a novel form of eye movement analysis to study how humans use simple graphical representations of uncertainty to solve an inference problem. Specifically, we examine how humans make a Bayesian inference using the frequency format, described in the previous section, represented with two different graphical layouts. We extract the strategies people use when solving a probabilistic reasoning problem by analysing the way they approach the problem. To determine how people arrive at a correct or incorrect inference, we record and analyse the eye movements of people trying to solve the weather problem described above. The information is presented using either tree or Venn diagrams, which are simple, familiar ways of demonstrating relationships between sets that have been used extensively in previous research (Binder, Krauss, and Bruckmaier 2015; Khan et al. 2015; Reani et al. 2018), and which demonstrate the relationship between items of information in different ways. Tree diagrams support reasoning through directionality – there is an order in which the information is read – whilst Venn diagrams emphasize explicitly the nesting between sets (Reani et al. 2018). The two visualizations display implicit and explicit aspects of the problem in different ways. Whilst Venn diagrams display more focused information about conjunctions between the sets, Tree diagrams present more comprehensive information in a linear fashion. Both the amount of information available, and the way in which it is presented, may affect reasoning. We analyze gaze data in combination with the answers participants provide to identify differences in reasoning strategy, and link them to the properties of the visualization. These representations also clearly separate in space the items of information required for reasoning, which is important within eye tracking studies. If a problem is in textual form, the items of information are not clearly separated, making it difficult to map a person’s gaze onto the components of the problem.

To understand the reasoning process, we considered both which items of information participants fixate on, and, crucially, how they compare these items, which can be ascertained from how they transition between them. Transition frequencies were represented as probability distributions, and a permutation test was used to assess whether the difference, calculated using a distance metric (Davies et al. 2016; Reani et al. 2018; Kübler, Kasneci, and Rosenstiel 2014), between those making correct inferences and those making incorrect inferences was significantly greater than the difference between groups created at random.

We hypothesised that people who reason correctly about problems asking them to make a prediction based on past data encoded in frequencies exhibit a different
gaze strategy to people who do not. We also hypothesized that reasoning about the same problem presented with two comparable, but conceptually different graphs can lead to different reasoning strategies. By visualizing gaze transitions directly on the two graphs and using appropriate analyses we found evidence which supports both of these hypotheses, and demonstrated how visual behaviour can help us understand cognition when studying reasoning.

2. Material and Methods

The experiment used a within-subjects design with one factor, Information Format. This had two levels, Tree and Venn. Correctness, the metric measuring participants performance, had two values, correct and incorrect. We adopted the strict correctness criteria proposed by Gigerenzer and Hoffrage (1995) as this has been widely used in previously published research to evaluate participants' responses to Bayesian reasoning problems similar to the one used in the current study (McDowell and Jacobs 2017). According to this criteria, only answers approximated correctly to the first non-decimal value were classified as correct answers.

2.1. Participants

Forty-nine participants (age range 16-36 years; 37 male and 12 female), were recruited from the University of Manchester to take part in a laboratory experiment conducted in university facilities. They were a convenience sample of students who self-enrolled in the experiment (which was advertised across campus) in exchange for an Amazon voucher. As numerical ability can affect reasoning capacity, a questionnaire was used to assess individual differences in numeracy (see next section for details). Participants were then presented with an adaptation of the weather problem on a computer screen while their eye movements were recorded.

2.2. Stimuli and Procedure

Participants sat in front of a computer and, after informed consent had been taken, they were presented with instructions on a computer screen explaining the nature of the study. They were also asked to complete a questionnaire to assess their numeracy, measured using the Subjective Numeracy Scale – ability subscale (Fagerlin et al. 2007). Participants were told that they would be shown two different hypothetical scenarios describing historical weather data for a fictional city. They were then presented with the following version of the weather problem adapted from Brase (2014) using the frequency format:

In Gotham city on average, out of 1000 days, some are rainy days. Of these rainy days, some are associated with low barometric pressure. Some of the remaining sunny days are also associated with low barometric pressure.

The problem was presented as text first, without data; then a second screen showed either a Tree or a Venn representation of the problem with the data. Specifically, two scenarios were presented sequentially, both referring to the same city. The presentation was self-paced and the participants could move to the second scenario only after
answering the question related to the first one. The presentation format and the frequency data for the past 1000 days differed (see Figure 1) – i.e., in one condition the data were presented using a Tree diagram in which the number of rainy days with low-pressure were 500 and the number of sunny days with low-pressure were 50; in the second condition a Venn diagram was used, where the figures were 350 and 100 days respectively. As we changed the frequency data in the two problems, we explicitly asked participants to imagine two different scenarios for the same fictional city. Previous research shows that changing the numerical values of the problem does not affect reasoning strategies, and we thus made the assumption it would not affect our analysis (Hafenbrd and Hoffrage 2015). The presentation of the scenarios was counterbalanced between participants to reduce carryover effects. For both tasks the participants were asked to predict the weather for the next 1000 days in Gotham City, based on the data about previous weather conditions provided in each scenario. They were asked, verbally, to fill in the blanks of the statement:

Out of a total of 1000 days, Gotham city has _____ days of low barometric pressure, _____ of which will be associated with it actually raining.

This answer format, taken from Brase (2014), was the same for both scenarios and was presented below the graph showing the associated data. This version of the problem, compared to problems which require an answer in the form of a single number, limits the numbers of possible non-Bayesian strategies that can be used, as the question explicitly asks participants to provide the two pieces of data which, in the probability version of the problem, would correspond to the fraction \( \frac{P(E \& H)}{P(E)} \). This format has been shown to increase the number of correct answers (Brase 2008, 2014; Girotto and Gonzalez 2001), which was desired in the current study. If only a small proportion of people answered correctly, as tends to happen with the standard probability format (Eddy 1982), there would be insufficient data to compare participants’ gaze behaviour between groups. The two diagrams presented two comparable scenarios which differed in the frequency data they displayed. Furthermore, it should be noted that the two graphs presented a different number of items of information, the consequences of which are considered in the Results and Discussion sections.

Participants were asked to answer the question below the graph verbally, providing the two figures: the number of days with low pressure (the sum of rainy and low pressure days, and sunny and low pressure days); the number of rainy days with low pressure. As the presentation of the two stimuli was counterbalanced, half of the participants were shown the Tree first and then the Venn representation, and vice versa for the other half. Responses were audio recorded. A Tobii X2-60 Eye Tracker, sampling rate 60 Hz, with Tobii Studio 3.2 software was used to record eye movements which were separately analysed using the R programming language. The analysis scripts and the data are available on GitHub (Reani 2018).

During reading tasks, fixation durations can generally vary between 100 and 800 milliseconds (ms), whereas saccade durations typically vary from 10 to 20 ms, although these can last up to 80 ms in certain cases (Raney, Campbell, and Bovee 2014). Thus, we set the fixation classification threshold to 80 ms, so that only events lasting longer than this threshold were classified as fixations (Holmqvist et al. 2011).
2.3. Analysis

A Mann-Whitney U test was used to determine whether Numeracy had any effect on Correctness. A multilevel logistic regression was used to determine whether there was a relationship between the predictor Format and the response variable Correctness, adjusting for the effect of the covariate Numeracy.

Descriptive gaze data statistics were calculated for dwell time and fixation frequency for the given areas of interest (AOIs) defined over the stimuli, as shown in figure 2. Transitions were defined as sub-sequences of 2 items (AOIs) in specific orders (e.g. R-R is the transition from total rainy days to rainy days with low pressure). An analysis of longer subsequences was not performed as the possible number of unique transition patterns increases dramatically with the pattern length, and each tends to occur with a much lower frequency, leading to sparse data that is difficult to interpret (Kübler et al. 2017; Reani, Peek, and Caroline 2018). To identify significant differences in the distribution of transitions between the Correct and Incorrect groups, two permutation tests, one for each condition, with 10,000 permutations each were performed using the Hellinger distance as defined by equation 2.

\[
H(P, Q) = \sqrt{\frac{1}{2} \sum_{x \in X} (\sqrt{p(x)} - \sqrt{q(x)})^2}
\]  

(2)

In Equation 2, \( p \) and \( q \) are the two discrete distributions to be compared over a vector of values \( x \), which here represents a vector of transitions with their relative frequencies (Hellinger 1909). This method was used as some of the assumptions for standard statistical tests were violated, it is robust against Type 1 error (Wilcox 2010) and it is suitable for comparing unevenly-sized groups that often occur when participants are assigned to these on a post hoc basis (Reani, Peek, and Caroline 2018). Odds-ratios are reported in the results section. These are used to find differences in the relative frequencies of transitions (or fixations in the case of single AOIs) between groups. The general principle is that the odds-ratio scale is used to find the transitions with the largest relative frequency, and which are thus responsible for the largest difference in the distribution of transition frequency between two groups (e.g. correct vs. incorrect respondents). Odds-ratios values closer to 1 represent similarities in transition relative frequencies (between groups), whereas values larger or smaller than one represent differences in transition relative frequencies between the compared groups. Details of how these were calculated are given below.
3. Results

All participants completed both problems. Of the 49 participants, 30 answered correctly with the tree format, and 29 with the Venn format, resulting in a correct response rate of approximately 60%. A MannWhitney U test indicates that Numeracy was higher for the Correct group (Mdn = 4.5) than for the Incorrect group (Mdn = 3.75), U = 737.5, p-value = 0.003. A regression analysis, controlling for the differences in participants’ numeracy, showed that information format (tree or Venn diagram) did not affect correctness - odds-ratio = 0.813, 95% CI [0.26, 2.54].

To perform the eye movement analysis, AOIs were defined over each stimulus (see figure 2), and coded using representative letters, such that gaze metrics could be calculated (see table 1).

<table>
<thead>
<tr>
<th>Code (Tree)</th>
<th>AOIs (Tree)</th>
<th>Code (Venn)</th>
<th>AOIs (Venn)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>Total number of days</td>
<td>T</td>
<td>Total number of days</td>
</tr>
<tr>
<td>Q</td>
<td>Question</td>
<td>Q</td>
<td>Question</td>
</tr>
<tr>
<td>Rh</td>
<td>Rainy &amp; high pressure</td>
<td>Rh</td>
<td>Rainy &amp; high pressure</td>
</tr>
<tr>
<td>Rl</td>
<td>Rainy &amp; low pressure</td>
<td>Rl</td>
<td>Rainy &amp; low pressure</td>
</tr>
<tr>
<td>Sl</td>
<td>Sunny &amp; low pressure</td>
<td>Sl</td>
<td>Sunny &amp; low pressure</td>
</tr>
<tr>
<td>Sh</td>
<td>Sunny &amp; high pressure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>Rainy days</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>Sunny days</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Tree diagrams display all the information found in Venn diagrams plus the marginal values (total rainy and total sunny days), and one additional intersection (Sunny Days & High Pressure Days), although it should be noted these additional pieces of information are not necessary to reach a correct inference. To answer the given question correctly, the reasoner must report the total number of days with low pressure (computed by summing the number of sunny days with low barometric pressure and the number of rainy days with low barometric pressure) and the number of rainy days with low barometric pressure. The items of information required to compute these values are found in both graphs, and correspond to Rl (Rainy-low) and Sl (Sunny-low) in table 1.
Figure 3. Proportional dwell time for each AOI, by group and by condition in percentages.

3.1. Dwell time

Dwell time per AOI, measured as the proportion of time spent fixating an AOI in relation to the time spent on all the AOIs together, varied between the Correct and Incorrect groups - see figure 3 and table 2, reporting medians and interquartile ranges for dwell time in percentages.

When using the Tree format, people in the Correct group spent more time fixating in locations Rainy, Sunny, Rainy-high, Sunny-low, Sunny-high and Question, and less time fixating in location Rainy-low than those in the Incorrect group. When information was displayed in the Venn format, people in the Correct group spent more time fixating in locations Sunny-low and Question, and less time fixating in location Rainy-low than those in the Incorrect group.

Table 2. Medians and interquartile ranges for proportion of dwell time (as a percentage) for each problem component in each group

<table>
<thead>
<tr>
<th>AOIs</th>
<th>TREE</th>
<th>VENN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correct</td>
<td>Incorrect</td>
</tr>
<tr>
<td></td>
<td>Mdn (%)</td>
<td>IQR</td>
</tr>
<tr>
<td>T</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Q</td>
<td>27</td>
<td>42</td>
</tr>
<tr>
<td>Rh</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>Rl</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>Sl</td>
<td>14</td>
<td>13</td>
</tr>
<tr>
<td>Sh</td>
<td>17</td>
<td>14</td>
</tr>
<tr>
<td>R</td>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td>S</td>
<td>6</td>
<td>11</td>
</tr>
</tbody>
</table>

3.2. Fixation frequency

When viewing information using the Tree format, the average number of fixations per participant was higher in the Correct group (m = 50.3, sd = 30.83), than the Incorrect group (m = 40.53, sd = 19.8). This was also true for the Venn format, where for the Correct group, m = 55.93, sd = 30.29, and for the Incorrect group, m = 50.58, sd = 23.07. For analysis purposes, relative frequencies were calculated by dividing the
frequency for a specific AOI by the total number of fixations across all the AOIs in that format. Table 3 shows the relative frequencies for each AOI according to format (tree/Venn) and group (Correct/Incorrect). The table also reports the odds-ratios measured as \( OR = (p / 1 - p) / (q / 1 - q) \), where \( p \) and \( q \) are the distributions of fixations across the AOIs of the two groups (Correct vs. Incorrect).

Table 3. Relative Frequencies of Fixations across AOIs, for Tree (left) and Venn (right)

<table>
<thead>
<tr>
<th>AOI</th>
<th>Correct</th>
<th>Incorrect</th>
<th>OR</th>
<th>AOI</th>
<th>Correct</th>
<th>Incorrect</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>0.05</td>
<td>0.08</td>
<td>0.62</td>
<td>T</td>
<td>0.06</td>
<td>0.08</td>
<td>0.69</td>
</tr>
<tr>
<td>Q</td>
<td>0.26</td>
<td>0.32</td>
<td>0.76</td>
<td>Q</td>
<td>0.32</td>
<td>0.23</td>
<td>1.58</td>
</tr>
<tr>
<td>Rh</td>
<td>0.08</td>
<td>0.09</td>
<td>0.92</td>
<td>Rh</td>
<td>0.2</td>
<td>0.19</td>
<td>1.04</td>
</tr>
<tr>
<td>Rl</td>
<td>0.1</td>
<td>0.08</td>
<td>1.32</td>
<td>Rl</td>
<td>0.15</td>
<td>0.27</td>
<td>0.46</td>
</tr>
<tr>
<td>Sl</td>
<td>0.13</td>
<td>0.14</td>
<td>0.92</td>
<td>Sl</td>
<td>0.28</td>
<td>0.23</td>
<td>1.33</td>
</tr>
<tr>
<td>Sh</td>
<td>0.19</td>
<td>0.18</td>
<td>1.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>0.09</td>
<td>0.04</td>
<td>2.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>0.1</td>
<td>0.08</td>
<td>1.19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The odds-ratios show that when viewing information in the tree format, the Correct group fixated more frequently in locations Rainy and Rainy-low and less frequently in locations Total and Question compared with the Incorrect group. Similarly, with the Venn format, the Correct group fixated more frequently in locations Sunny-low and Question and less frequently in locations Rainy-low and Total compared with the Incorrect group.

### 3.3. Fixation Duration

For both groups and both formats, the average fixation duration was 197ms (sd = 139ms). The distribution of these data was positively skewed - i.e., shorter fixations occurred more frequently than longer ones. There were nonetheless 150 long fixations (about 5%) which exceeded the duration of half second.

Table 4. Medians and quartiles of the distribution of fixation duration across conditions and groups

<table>
<thead>
<tr>
<th>Condition</th>
<th>Group</th>
<th>Q1</th>
<th>Mdn</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tree</td>
<td>Correct</td>
<td>117</td>
<td>167</td>
<td>217</td>
</tr>
<tr>
<td>Tree</td>
<td>Incorrect</td>
<td>102</td>
<td>150</td>
<td>217</td>
</tr>
<tr>
<td>Venn</td>
<td>Correct</td>
<td>133</td>
<td>166</td>
<td>233</td>
</tr>
<tr>
<td>Venn</td>
<td>Incorrect</td>
<td>117</td>
<td>167</td>
<td>217</td>
</tr>
</tbody>
</table>

In table 4 the medians and quartiles of the distributions of fixation duration across conditions and groups are reported (in milliseconds). Some fixation durations extended well over the distribution boundary (for outliers) of 389.5ms, which is the largest boundary among the four groups – this boundary is computed as \( B = Q3 + 1.5IQR \). Previous research shows that longer fixations indicate that participants are engaged in greater information processing, or are experiencing higher cognitive load (Debue and van de Leemput 2014). As 217ms is the 3rd quartile for all but the Venn/Correct group, this is used as the threshold for a long fixation. Table 5 shows the relative frequencies of long fixations (those that exceed the duration of 217ms) and their related odds ratios across both groups and formats.
Table 5. Relative Frequencies of Long Fixations across AOIs, for Tree (left) and Venn (right)

<table>
<thead>
<tr>
<th>AOI</th>
<th>Correct</th>
<th>Incorrect</th>
<th>OR</th>
<th>AOI</th>
<th>Correct</th>
<th>Incorrect</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>0.05</td>
<td>0.07</td>
<td>0.71</td>
<td>T</td>
<td>0.03</td>
<td>0.07</td>
<td>0.45</td>
</tr>
<tr>
<td>Q</td>
<td>0.24</td>
<td>0.34</td>
<td>0.63</td>
<td>Q</td>
<td>0.28</td>
<td>0.23</td>
<td>1.31</td>
</tr>
<tr>
<td>Rh</td>
<td>0.07</td>
<td>0.06</td>
<td>1.27</td>
<td>Rh</td>
<td>0.22</td>
<td>0.17</td>
<td>1.42</td>
</tr>
<tr>
<td>Rl</td>
<td>0.1</td>
<td>0.13</td>
<td>0.71</td>
<td>Rl</td>
<td>0.14</td>
<td>0.25</td>
<td>0.48</td>
</tr>
<tr>
<td>Sl</td>
<td>0.11</td>
<td>0.12</td>
<td>0.92</td>
<td>Sl</td>
<td>0.32</td>
<td>0.28</td>
<td>1.24</td>
</tr>
<tr>
<td>Sh</td>
<td>0.21</td>
<td>0.17</td>
<td>1.23</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>0.13</td>
<td>0.06</td>
<td>2.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>0.09</td>
<td>0.05</td>
<td>2.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

When using the Tree format, the Correct group had a higher relative frequency of long fixations in locations Rainy and Sunny, and a lower relative frequency of long fixations in locations Total, Rainy-low and Question compared with the Incorrect group. For the Venn format, the Correct group had a higher relative frequency of long fixations in locations Rainy-high, Sunny-low and Question, and a lower relative frequency of long fixations in locations Rainy-low, and Total, compared with the Incorrect group.

3.4. Transition Analyses

In this analysis we did not consider within AOI transitions, as we are primarily interested in how people compare items of information. The Tree format had 8 AOIs, yielding 56 potential transitions. The Venn format had 5 AOIs, yielding 20 potential transitions. A permutation test compared the Hellinger distance between the Correct and Incorrect groups against the distance between two groups created at random. The estimated sampling distributions of the two tests are shown in figure 4. The vertical purple line represents the distance between correct and incorrect groups for the tree and Venn conditions respectively. The grey curve represents the distributions of the distances between pairs of randomly sampled groups (with replacement) of comparable sizes. The cut-off area under the curve delimited to the right of the vertical line is the probability of the null hypothesis being true; i.e., that the distance between the transition distributions of correct and incorrect group is not different from the distances between any two groups of comparable sizes sampled at random from the population.

For the Tree format, the Hellinger distance between the distribution of transitions of the correct group and the distribution of transitions of the incorrect group was $H_d = 0.315$; the p-value was $p = 0.022$. For the Venn format, the Hellinger distance between the distribution of transitions of the correct group and the distribution of transitions of the incorrect group was $H_d = 0.196$; the p-value was $p= 0.161$. These results indicate that for the Tree format in particular, there are transitions that discriminate between groups. The transitions that best discriminate gaze behaviour between the Correct and Incorrect groups were identified based on the differences in relative frequency between the two groups.

To determine whether a transition is a meaningful discriminator, two parameters were taken into account. The first one is the transition odds-ratio value. For this, the odds-ratio scale was computed (following transformation using add-one Laplace smoothing) as $OR = (p / 1 – p) / (q / 1 – q)$ where $p$ (for Correct) and $q$ (for Incorrect) are the distributions of transitions of each of the groups. An odds-ratio
Figure 4. Sampling distributions of the Hellinger distances. The vertical red line represents the distance between correct and incorrect group.

Figure 5. Transition distribution by odds-ratio (x-axes) and absolute frequency (y-axes) for Venn (right) and Tree (left). The graph axes were log-transformed. The red dots are the only transitions with the odds-ratio confidence interval (CI) which does not include 1. The vertical continuous line represents odds-ratio = 1. The vertical dashed lines and the horizontal dashed line are the limits within which odds-ratio CIs include 1.

value close to one indicates that that transition is found in both groups with a similar relative frequency, and is thus not a useful discriminator. The second parameter is the maximum absolute frequency of the transitions, computed as $F = \max(x_i, y_i)$ the maximum value between the transition frequency for the correct group and the transition frequency for the incorrect group. Thus, $F$ represents the absolute transition frequency for one group, the one with the largest value. Transitions that have odds-ratios very far from one, but that occur only few times, are not reliable indicators of the strategies used by a typical reasoner, as they are made by only a few participants. These two dimensions are shown for both conditions in figure 5 which represents the transition distribution by odds-ratio (x-axis) and absolute frequency (y-axis) for Venn (right) and Tree (left). The red dots are the transitions with the odds-ratio confidence intervals (CI) that do not include 1. The vertical continuous line represents an odds-ratio of value one. The vertical dashed lines and the horizontal dashed line are the limits within which odds-ratio CIs include one. These lines are computed by taking the extremes of $F$ and OR for the transitions for which the CI did not include the value of one. Ideally, one should look for transitions that are delimited by the dashed lines in the areas found in to the top left and the top right of the graphs; these transitions have a high frequency and are also very discriminative (i.e., far from OR = 1).

As figure 5 shows, discriminative transitions with a high frequency are relatively
Figure 6. Discriminative eye-movements in the tree format (top) and in the Venn format (bottom), for the Correct (left) and Incorrect (right) groups.

rare. For the Tree condition, the two discriminative transitions with a small odds-ratio value found on the top left area isolated by the dotted lines of figure 5 were transition Sl-T (from Sunny-low to Total), OR = 0.08, 95% CI [0.01, 0.64], F = 6; and transition S-T (from Sunny to Total), OR = 0.28, 95% CI [0.08, 0.93], F = 7. For the Venn condition, the two discriminative transitions with a small odds-ratio value were Rh-Rl (from Rainy-high to Rainy-low), OR = 0.57, 95% CI [0.34, 0.95], F = 36; and Sl-T (from Sunny-low to Total), OR = 0.39, 95% CI [0.15, 0.99], F = 11. Also for in the Venn condition, there were two discriminative transitions with a large odds-ratio value: Sl-Q (from Sunny-low to Question), OR = 4.33, 95% CI [1.89, 9.93], F = 33; and Q-Sl (from Question to Sunny-low), OR = 2.55, 95% CI [1.30, 4.99], F = 36.

To demonstrate how participants gaze behaviour was linked to reasoning strategies, the discriminative transitions described above are plotted on top of the tree and Venn diagrams in figure 6. The dotted arrows represent the transitions that have large or small odds-ratios, of a high frequency, and a narrow confidence interval that does not include the value of one. Grey shading of the AOI represents a longer dwell time, and darker borders represents larger number of fixations (for that group compared to the other). The most apparent effect shown in figure 6 is that, in the incorrect group, there is a tendency for participants gaze to transition to location Total (the total
number of days). A second observation is that the visual attention of participants who answered correctly shifts between location Sl (sunny low) and Q (the question). The interpretation of these two phenomena is fully addressed in the discussion.

3.5. Analysis of participants’ answers

The question required participants to make a prediction based on historical data. This is accomplished by selecting and combining the two quantities representing the fraction \( \frac{R_l}{(R_l + S_l)} \), where \( R_l \) is Rainy-low – i.e., \( P(E \& H) \), and \( S_l \) is Sunny-low – i.e., \( P(E \& H^\prime) \). Of 98 responses, 39 were incorrect (39.8%). We analysed these incorrect responses and grouped them by their similarity (Gigerenzer and Hoffrage 1995), to determine commonly used strategies. Figure 7 shows a map of participants responses representing these strategies. We distinguish here totally incorrect answers, for which both numbers (the numerator and the denominator) in the answer were incorrect, from partially incorrect answers, where at least one number was correct. In the Tree condition there were 19 incorrect responses of which 7 were totally incorrect; in the Venn condition there were 20 incorrect responses, of which 11 were totally incorrect. Four of these had 1000 days (the AOI Total) as the denominator of the fraction and different values \( x \) in the graph for the numerator. In the Tree condition, two of the 12 partially correct answers had the correct numerator \( R_l \) (Rainy-low), but used \( T \) (the Total) as the denominator; 3 responses had the correct numerator but had \( S_l \) (sunny-low) as the denominator; 1 response had the correct numerator but had \( R_h \) (Rainy-high) as the denominator; 6 responses had the incorrect numerator \( R \) (Rainy) but the correct denominator \( R_l + S_l \) (Rainy-low plus Sunny-low). For the Venn condition, the 9 partially correct answers were all the same, and had the correct numerator, but the incorrect denominator \( S_l \) (sunny low). These results suggest that there was consistency in certain reasoning strategies across participants who answered incorrectly to the question, and that these varied by the format in which they were presented. The explanation of these finding is fully addressed in the following section.

4. Discussion

In this research, we demonstrate how eye tracking methods can be used to study how visual presentation of information affects human reasoning. We used a Bayesian problem with two different graphical representations and investigated whether the type of representation influences people’s reasoning strategies by analysing participant gaze behaviour and correlating this with their responses. The study suggests that the layout of the information does indeed influence the strategy people use. In the following discussion we report the evidence from different measures that supports this conclusion.

4.1. Dwell Time, Fixation Frequency and Duration

The stimuli components that are relevant for solving the weather problem are \( R_l \) (Rainy-low) and \( S_l \) (Sunny-low). To study participants reasoning, we examined the extent to which people attend to these and other components of the graph during problem completion.

Firstly, it is worth noting that participants in the correct group spent on average
Figure 7. Strategies used by participants classified according to the group and condition to which they belong. The correctly identified quantities are in bold font.

more time on the question and, in the Venn condition, they also fixated on the question more frequently. In general, paying more attention to the question appears to be associated with better performance. In the tree condition, correct respondents spent less time and fixated more frequently on RI and spent more time and fixated less frequently on SI. In the Venn condition, correct respondents looked for a shorter time and less frequently at location RI and for longer time and more frequently at location SI. This indicates that the layout of the graphs affects how people perceive the information. The nested character of Venn diagrams emphasizes the conjunction RI (Rainy AND low pressure); this AOI therefore receives less attention when people understand the problem presented. Peripheral data in Venn diagrams, however, is not explicitly displayed using nesting; thus, this may be the reason why correct respondents spent more time and looked more often at location SI (Sunny-low), for which the conjunction was less apparent. These results draw a parallel with findings from past research (Böcherer-Linder and Eichler 2017) which showed that graphs emphasizing nested-set structures give quicker access to the information needed for solving Bayesian reasoning problems when compared with tree diagrams, which show the information in a hierarchical structure.

Another interesting observation is that, in the tree condition, correct respondents looked at location RI (Rainy-low), which was one of the two key information components required to answer the question correctly, for a shorter time but more frequently than incorrect respondents. This was explained by a larger number of long fixations in this location made by participants in the incorrect group. This behaviour was also seen in the Venn condition. As longer fixations are known to indicate higher cognitive load (Debue and van de Leemput 2014), this result suggests that participants who answered incorrectly found understanding what this piece of information referred to, in relation to the problem, difficult to process.

4.2. Gaze Transitions

The permutation tests show a link between transition behaviour and correctness, which was strongest in the tree condition. Figure 6 shows that participants in the incorrect group tended to look more often towards the total compared with participants in the correct group. This was found in both graphs suggesting that this behaviour was
independent of the graphical representation of the problem. We call this tendency Re-
version to Total. This behaviour is consistent with the fact that some of the people who
answered incorrectly reported the total number of days, rather than the subset of the
population of interest (see figure 7). This is also consistent with previous research in
which a large proportion of the participants who correctly identified the numerator of
the proportion required for the answer, failed to select the correct denominator, choos-
ing the entire population instead (Khan et al. 2015; Reani et al. 2018). Figure 6 also
shows that participants in the correct group, when using the Venn diagram, tended to
check the question more often. This behaviour was not found when the problem was
shown using the tree diagram, indicating it was format dependent. The Tree and Venn
diagrams used in the study are different in a number of ways. Firstly, tree diagrams
include more information. This difference alone could have directly affected people’s
strategies. For instance, trees may be more self-explanatory compared to Venn as they
show the complete set of intersections. Secondly, Venn diagrams have an intersection
of different category types (e.g., rainy days & low pressure days), whereas tree dia-
agrams have the categories of rainy/sunny days and low/high pressure days displayed
in different levels of organization. Due to the way the Venn diagram is constructed,
the intersection between the sets Sunny AND High pressure days is not explicitly la-
belled. This is the complement of the union of the events Rainy and Low pressure
days; i.e., the area outside the two circles. As a consequence of these differences, the
relationships between the numbers in each diagram are different. Specifically, the tree
structure is neutral - i.e., it shows both subsets, rainy and sunny, with both levels,
low and high pressure, in a parallel manner), whereas the Venn diagram with category
circles is chosen to serve the specific purpose of the particular task question - i.e., it
shows only selected information, omitting the subsets that are not necessary to answer
the question (e.g., sunny-high).

4.3. Answers Analysis

Both visualizations appear to have advantages and disadvantages. In this section we
report these in relation to the problem at hand, linking the results with the previous
literature (McDowell and Jacobs 2017; Khan et al. 2015; Gigerenzer and Hoffrage
1995).

One might expect different gaze patterns in graphs with different visual layouts.
The question is, do they provide evidence of different reasoning strategies? When the
information was presented using a tree diagram, the most common strategy used by
participants who answered incorrectly was to choose the number of rainy days (rather
than the number of rainy days with low pressure) as the numerator of the fraction
(i.e., the second quantity required in the question). It appears that the participants
who used this strategy failed to understand the conjunction between sets (Rainy AND
Low). This strategy was not found once in the Venn condition and may be linked to
the fact that Tree diagrams explicitly display the quantity Rainy days, whereas Venn
diagrams do not. To extract this quantity in Venn diagrams, one needs to make an
algebraic calculation - i.e., summing Rl (Rainy-low) to Rh (Rainy-high). This extra
step in the computation of this quantity may protect participants against adopting
this faulty strategy. This shows one advantage of using a more focused graph over a
graph which display the full information set.

When the problem was presented using a Venn diagram, the most common strategy
used by participants who answered incorrectly was to replace the correct denominator
i.e. Rl (Rainy-low) plus Sl (Sunny-low) - with Sl alone. This strategy is what Gigerenzer and Hoffrage (1995) called the comparison shortcut, which may lead to a close-to-correct answer only if Sl were much larger than Rl. However, in the problem presented in this study this was not the case and this strategy produced a result that differed considerably from the correct answer. Although, this strategy was also found in the tree condition, it was less common for this visualization. It may be that tree diagrams shield the reasoner from the comparison shortcut as they logically show how the leaves in the tree are derived using a directional process. This is an advantage of a more comprehensive graph that shows the full information set.

Participants made a greater variety of errors when using the Tree diagram than they did when using the Venn diagram. This may be due to the fact that the Tree representation presents more items of information, and thus provides the opportunity for more diverse types of faulty reasoning.

In summary, Venn diagrams show less and more focused information. This enables participants to easily identify the conjunction between the rainy days and low pressure days sets. Tree diagrams support the reasoner in identifying the marginal information (i.e., low pressure days), due to the way the leaves of the diagram unfold from their branches. This produces a logical flow linked to the way the diagram is read from left to right. As a result, if the intent of the visualization is to show conjunctions between sets, Venn diagrams should be preferred; whereas, if the intent of the graph is to enable users to derive the marginal data, then, Tree diagrams should be used instead. In the weather prediction problem used in our study, participants were required to identify both pieces of information. It follows that neither of the visualizations was strictly dominant as they both have advantages and disadvantages for supporting different aspects of the process for correctly reasoning about this.

4.4. Conclusion

Our study provides evidence that different types of visualizations prompt people to reason in different ways. Evidence for this is found in participants’ answers, and is supported by data showing differences in gaze behaviour. We also show that gaze behaviour is linked to performance, as people who answered correctly displayed a different distribution of transitions and fixations to those who answered incorrectly. This work deepens our understanding of how human interact with graphical representations of uncertainty and has implications for the study and practice of information visualization design.

One of the limitations of our study is that individual differences, other than numeracy, could have affected the results. In cognitive experiments, it is not uncommon to recruit self-volunteering students. However, we acknowledge that sampling participants using refined criteria from different populations would improve the generalizability of our findings. It is worth noting that in the current study we adopted a within-subjects design to assess the effect that the type of visual representation has on human reasoning, which was the main variable of interest of the study, and counterbalanced the conditions of this variable. This design choice greatly reduced this hypothetical confounder, thus providing strong internal validity.

The second limitation of our work resides in its reliance on eye-tracking methods to study the reasoning processes. Eye-tracking data is a proxy for cognition. In general, where people look, for how long, and in what order are useful indicators of what may be going on in a persons mind, but there is not necessarily a direct link. A benefit of
eye-tracking in the study of human reasoning, is that it provides an objective form of evidence that does not influence the reasoning process. This cannot be guaranteed with methods such as think aloud or stimulated recall techniques (Blondon, Wipfli, and Lovis 2015). Combining this data with an understanding of the graph, and the way people answer the problem, contributes further to our understanding of how human reasoning unfolds over time. There are other graphs that can be used to display uncertainty which were not investigated in the current research. Thus, future work will focus on comparing different graphs, and collecting other forms of interaction data such as mouse movements and clicks, to further investigate the relationship between the graphical representation of uncertainty, and the way people reason about it.

**Declarations of interest:** None

**References**


Cosmides, Leda, and John Tooby. 1996. “Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judg-


Holzinger, Andreas, Markus Plass, Katharina Holzinger, Gloria Cerasela Crișan, Camelia-M. Pintea, and Vasile Palade. 2016. “Towards interactive Machine Learning (iML): Applying Ant Colony Algorithms to Solve the Traveling Salesman Problem with the Human-in-the-


Highlights:

- We studied probabilistic reasoning in the context of human-graph interaction.
- We proposed an eye tracking method to study human reasoning.
- Probabilistic reasoning is affected by how uncertainty is graphically encoded.
- Gaze behaviour is linked to people performance in Bayesian reasoning tasks.