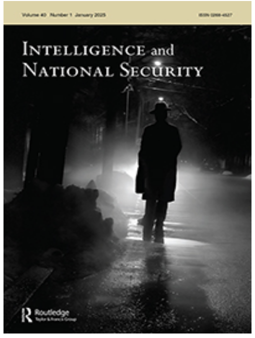

Intelligence and National Security

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/fint20

Visualizing versus verbalizing uncertainty in intelligence analysis

Mandeep K. Dhami, Jessica K. Witt & Peter De Werd

To cite this article: Mandeep K. Dhami, Jessica K. Witt & Peter De Werd (15 Mar 2025): Visualizing versus verbalizing uncertainty in intelligence analysis, Intelligence and National Security, DOI: 10.1080/02684527.2025.2468049

To link to this article: https://doi.org/10.1080/02684527.2025.2468049

Published online: 15 Mar 2025.

INTELLIGENCE AND NATIONAL SECURITY
https://doi.org/10.1080/02684527.2025.2468049

Routledge
Taylor & Francis Group

RESEARCH ARTICLE

Visualizing versus verbalizing uncertainty in intelligence analysis

Mandeep K. Dhami, Jessica K. Witt and Peter De Werd

ABSTRACT

We compared the probability terms used by Western intelligence organizations against two visual encoding channel-based representations of uncertainty (i.e. darkness and thickness). Analysts were more sensitive to the probability being communicated under the word than thickness condition but not the darkness condition, with no difference among the visual conditions. However, sensitivity was not perfect. There was no difference in inter-individual variability across all conditions, which was generally poor. Test-retest reliability was greater in the word compared to thickness condition, but not the darkness condition, although, it was imperfect. Finally, analysts did not fully comply with existing uncertainty communication lexicons.

ARTICLE HISTORY
Received 4 January 2024
Accepted 11 February 2025

Introduction

Policies and decisions in many consequential domains, including climate science, law, medicine, and defense and security, are often based on 'expert' subjective probability judgments, which are made under conditions of uncertainty.[1] In the intelligence analysis domain examined in the present study, even well-informed and reasoned judgments can lead to poor outcomes if the probabilities (or uncertainty) associated with them are misunderstood by users.[2] Indeed, intelligence failures associated with the miscommunication of uncertainty do occasionally occur,[3] and so intelligence organizations have implemented policies for communicating uncertainty at both national and international levels.[4]

Even though probability can be communicated using one or a combination of three different formats (i.e., words, numbers, and visualizations), Figure 1 shows that Western intelligence communities recommend using words or verbal probabilities.[5] Their 'standardized lexicons' typically comprise a small set of terms ordered from the lowest to highest level of probability; and often with each term associated with a numerical range.[6] However, these policies run counter to psychological evidence which suggests that verbal probabilities may not be the most effective format for communicating uncertainty unambiguously.[7]

Given that the intelligence community has eschewed the use of numbers as a primary means of communicating uncertainty,[8] in the present study we therefore explore the utility of an alternative approach, namely the use of visual representations of uncertainty. Recent work adds to a small but growing body of literature recognizing the value of visualizations for time-constrained and overwhelmed intelligence consumers.[9] To inform a broader audience of policymakers and the public, some current risk assessment models incorporate colour-coding or infographics. Examples include the Department of Homeland Security's Advisory System[10] and the Dutch Terrorist Threat Assessment.[11] Similarly, analysts also draw on colour-coding systems to represent different risk levels or visualise an overview of the battlespace, facilitating situational awareness and the rapid

CONTACT Mandeep K. Dhami m.dhami@mdx.ac.uk

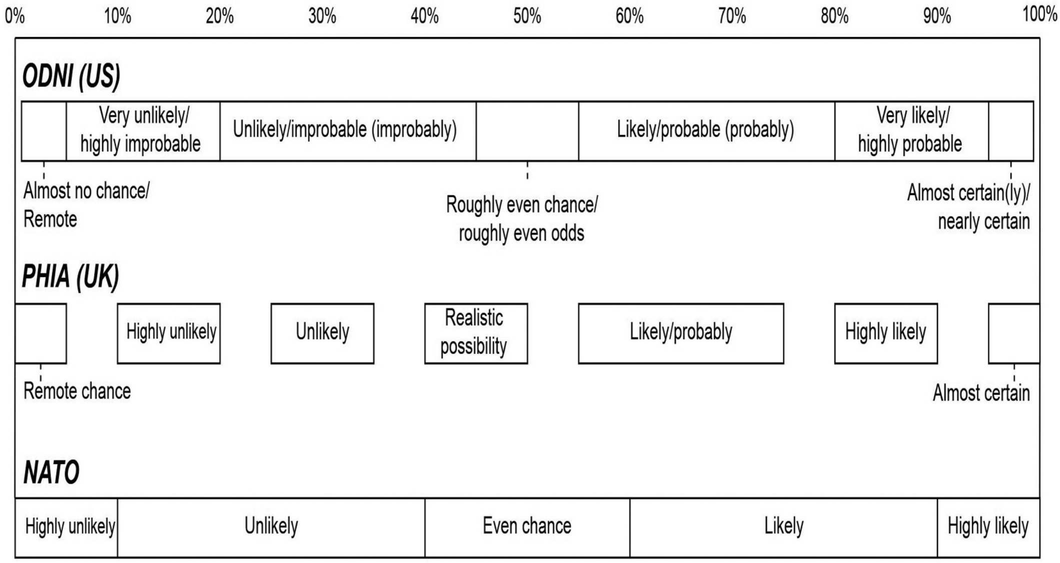

Figure 1. Lexicons for communicating probability in intelligence assessments (taken from Dhami and Mandel 2021).

comprehension of complex data.[12] Furthermore, proposals have been made to use visualizations such as fading colour to communicate uncertainty in the intelligence analysis domain,[13] with efforts already underway in this regard.[14] However, to-date, no-one has empirically examined whether such visualizations perform better than words in communicating uncertainty unambiguously, as we do. In this way, the present study connects the literature on verbal probabilities with that on visual representations of uncertainty. The study also contributes to the broader effort to use psychological research and methods to inform policy making in the intelligence analysis domain.[15] Before describing the aims of the present study along with the methods and results, we briefly review relevant past literature on communicating uncertainty.

Communicating uncertainty using verbal probabilities

People naturally prefer to communicate uncertainty verbally,[16] and this tendency has also been observed in the intelligence analysis domain.[17] According to Zimmer, the main advantages of using words to communicate uncertainty lies in the idea that people find it easier and more natural to use words.[18] Indeed, the language system is learned early during individual development, and words are a dominant mode of human communication. Therefore, verbal probabilities fare well in terms of ease of expression. Despite these advantages, research has pointed to the drawbacks of verbal probabilities as a means of communicating uncertainty unambiguously.

Perhaps most importantly, studies have repeatedly demonstrated that individuals have imprecise or vague numeric interpretations of verbal probabilities, and that there is inter-individual variability in interpretations of probability terms.[19] These findings have also been observed in the intelligence analysis context,[20] where an early study suggested that consumers of intelligence assessments had lower numeric interpretations of terms than did analysts who made the assessments.[21]

Efforts to mitigate against the imprecision and variability in interpretation of verbal probabilities by assigning numeric ranges to them (see Figure 1) have also been shown to be largely ineffective.[22] Research suggests that people cannot easily suppress their personal numeric meanings of probability terms and adopt new (mandated) meanings.[23] In addition, the language for communicating probability, even in the intelligence context, can ‘go-out-of-fashion’.[24] Studies have also documented a lack of compliance or agreement between users' (both senders' and receivers') numeric interpretations of terms and the intended or mandated meaning of these terms in the standardized lexicons currently in operation in the intelligence community.[25]

The aforementioned problems associated with verbal probabilities may be exacerbated in a collaborative analysis context since different organizations use different terms to represent the same probability, and the same term to represent different probabilities (see Figure 1). In addition, the effective communication of uncertainty using language-based formats may be affected by intelligence sharing across native and non-native (English-)language speaking countries such as in NATO; an issue that will be examined for the first time here. Past research on uncertainty communication in the intelligence analysis domain has focused on native English-language speaking countries, while research on uncertainty communication in other domains has translated English verbal probability expressions into other languages.[26] Thus, relatively little is known about how non-native English-language speakers numerically interpret English verbal expressions of uncertainty.

Finally, the use of verbal probabilities may be particularly problematic in the intelligence analysis domain given the need to communicate the likelihood of rare events, the need to integrate and update multiple probabilities associated with the same event, and the requirement to be policy neutral. Research suggests that despite efforts to select terms that may convey small probabilities of less than .1, these are interpreted as much greater by analysts,[27] thus resulting in over-estimates. It has also been found that individuals have difficulty in aggregating verbal probabilities,[28] and analysts use 'guesswork' in doing so, which results in lower accuracy.[29] Finally, probability terms may be perceived as providing implicit policy recommendations, even though these are not communicated explicitly,[30] thus undermining policy neutral analysis.

In light of the concerns associated with using verbal probabilities, some critics have argued for the use of numbers.[31] Unlike verbal probabilities which have unreliable ordinal scale properties, numbers have reliable ratio scale properties. This can therefore reduce imprecision and variability in interpretations. Numeric probability representations include percentages and decimals, and their values can be precise (e.g., 'there is a .65 chance') or imprecise (e.g., .55 to .70). However, as Dhami and Mandel point out, the intelligence community has historically been averse to using numeric probabilities.[32] This has been partly due to misconceptions about probability and its quantification[33] and partly due to the attitude that analysis is an 'art' and analysts are not 'mathematicians'.[34] Other arguments against the use of numeric probabilities more generally refer to the concern that they require reasonable levels of numeracy.[35] Given the apparent reluctance of the intelligence community to express probability using numbers primarily, in the present study, we explore the utility of an alternative approach, namely the use of visual representations of uncertainty.

Others have previously proposed the use of visualizations to communicate uncertainty in the intelligence analysis domain[36] and some organizations are pursuing this approach.[37] As Padilla et al.[38] point out, probability can be communicated visually by either graphical annotations of distributional properties of data as in error bars, icon arrays, violin plot and an ensemble plot[39] or via visual encoding channels such as colour, position and size.[4041] Graphic annotations of distributional properties of data are often used in communication of future risks,[42] and have been widely examined in psychological research on risk communication, particularly in the medical domain.[43] The present study, by contrast, focuses on visual encoding channel-based representations of uncertainty that are often used in geographic information systems and cartography,[44] and which are commonly employed in visual data analytics (e.g., geospatial analysis or social network analysis) used in the defense and security domain.[45]

Communicating uncertainty via visual encoding channels

Visualizations afford rapid processing of information,[46] which may be beneficial in time-sensitive (defense and security) environments. The fact that visualizations engage perceptual processes means that they are less likely than words or numbers to overload the cognitive system which is processing the data, and consequently there may be few speed-accuracy trade-offs.[47] For instance, visualization of the uncertainty in an adversary's current military capability could be processed at the perceptual level whereas information on what that capability was (e.g., number of tanks, self-propelled weapons, personnel etc.) could be processed at the cognitive level. Visualizations can also fairly easily communicate multiple probabilities simultaneously,[48] which is expedient when dealing with complex environments.

Visualizations can additionally be dynamic and interactive, allowing the user to contemplate data over time or under different situations.[49] Uncertainty information represented as darkness or thickness can either be integrated or superimposed on the data being visualized or presented separately (adjacent) to the data. When probability information is 'integrated' with the data it is less likely to be ignored or sub-optimally weighted in decision-making.[50] Finally, visual encoding channel-based representations of uncertainty, in particular, can overcome the problems associated with language barriers and numeracy that may mire the use of verbal or numeric probabilities, respectively.

However, despite their potential benefits, these visualizations do have potential drawbacks. Perhaps most obviously is the fact that unlike verbal probabilities, which are language-based alternatives to numeric expressions of uncertainty, visual encoding channel-based representations such as darkness and thickness (e.g., of links connecting entities to a node in a social network chart) are not expected to map onto numbers. Indeed, these visualizations have no intrinsic meaning and so are not explicitly scaled and consequently may suffer from imprecision and variability in interpretation to an even greater extent than do verbal probabilities. Relatedly, following the work of Hsee and colleagues on the 'evaluability hypothesis', one could argue that unlike with probability terms which can be rank ordered along (or map onto) the 0–1 probability interval,[51] levels of a visualization are more difficult to evaluate in this way, especially if they are presented separately (or consecutively) rather than jointly (simultaneously). Past studies have revealed that there is no universal agreement in how visualizations should be encoded or ordered from low to high probability, although a reasonably reliable degree of ordering can be achieved.[52] Users may thus need to be provided with instructions on how to encode an uncertainty visualization, and as with standardized verbal probability lexicons, such instructions typically comprise assigning numeric values (precise or imprecise) to each incremental step in the visualization.[53] Finally, people may not reliably and accurately perceive many different levels of a visualization[54] and so, as with words, this coarsens the probability scale, making the chances of a rare event difficult to communicate.

Few studies have directly compared uncertainty visualizations with verbal probabilities, and they have not compared the two formats in terms of their numeric interpretations, as we do. Hogan Carr et al. reported that participants in their focus groups preferred a combination of visual graphics and text (where the graphics provide a quick depiction of a risk and the text provides a concise explanation).[55] Participants wanted to avoid overly technical language and to have intuitive colour schemes, with different colours limited to no more than seven. Milne et al'.s survey respondents thought that compared to numeric and visual representations of uncertainty, verbal probabilities did not provide enough information and were less straightforward to interpret (i.e., less clear and needed more explanation).[56] Overall, recent large-scale reviews of uncertainty visualization research have lamented that few representations have been rigorously assessed.[57] Therefore, in an effort to contribute to the existing evidence-base for communicating uncertainty in intelligence analysis, the present study not only refers to the literature on verbal probabilities, but also refers to the literature on visual encoding channel-based representations of uncertainty.

The present study

The primary aim was to compare verbal and (two) visual uncertainty communications. Specifically, we compared verbal probabilities from the existing intelligence community lexicons against two visual encoding channel-based representations of uncertainty (i.e., darkness and thickness of links connecting entities to a node in a social network chart). These two visual formats have been previously found to be rated by users as intuitive[58] and have performed well in tests of accuracy of judgments based on the probability or uncertainty being presented.[59] In the present study, these two visual representations of uncertainty were integrated or intrinsic to the task participants were asked to perform (see Method section).

The verbal and visual formats were assessed on the following metrics of between- and within-individual performance, which were extrapolated from participants' probability estimates: (1) extent of agreement across individuals in probability encoding direction (e.g., do individuals agree that darker links represent greater probability?); (2) each individual's sensitivity to the probability being communicated (e.g., do individuals correctly grasp the probability that the word *very likely* is intended to convey?); (3) inter-individual variability in estimation; and (4) intra-individual variability in estimation (i.e., test-retest reliability).

The dearth of past research comparing verbal probabilities with the use of darkness or thickness to communicate uncertainty precludes a priori directional hypotheses. However, the fact that people commonly use words (rather than darkness or thickness) to express uncertainty may confer an advantage to the verbal format. Similarly, the idea that probability terms can be mapped onto the 0–1 probability interval whereas levels of darkness or thickness are more difficult to evaluate in this way suggests that the verbal format may outperform the visual format.

On the other hand, the advantage of the verbal format may be reduced for non-native English-language speakers, who are participants in the present study, because a secondary aim was to contribute to the small body of past research examining analysts' compliance with existing lexicons (i.e., the extent to which they interpret the terms as mandated). Whereas past research on this issue has been conducted with native English-language speakers,[60] we examine compliance among non-native English-language speakers from a NATO country (i.e., The Netherlands). We anticipated that compliance rates would be low for this population partly because Renooij and Witteman found that a Dutch-speaking sample expressed uncertainty using terms largely different from those in the existing lexicons.[61] In addition, compliance may be low because errors may arise in foreign language translation, and some probability terms may not exist in another language and so cannot be easily interpreted.

Method[62]

Participants

Sixty-two Dutch intelligence analysts and officer cadets attending training at the Defence Intelligence and Security Institute in the Netherlands or the Defence Academy volunteered to participate in the study without reimbursement. Participation was anonymous. Eighty percent of the sample was male, and the average age was 32.48 years (*SD* = 8.99, min = 18, max = 54).

Design

We used a 3 × 8within-subjects experimental design. The independent variables were Communication Format, which had three levels (i.e., word, darkness, thickness) and Probability Level, for which there were eight levels as we describe below.

Stimuli and measures

Participants were asked to complete three tasks that involved judging the likelihood of each of 12 individuals being a member of an insurgent group. Each task was presented as a network chart with the individuals labeled only as letters (e.g., 'A') and linked to an unlabeled central node. Hence, all 12 stimuli (links) were presented simultaneously. The length of the links and the distance between them was identical. The only difference across the three tasks was the format in which uncertainty

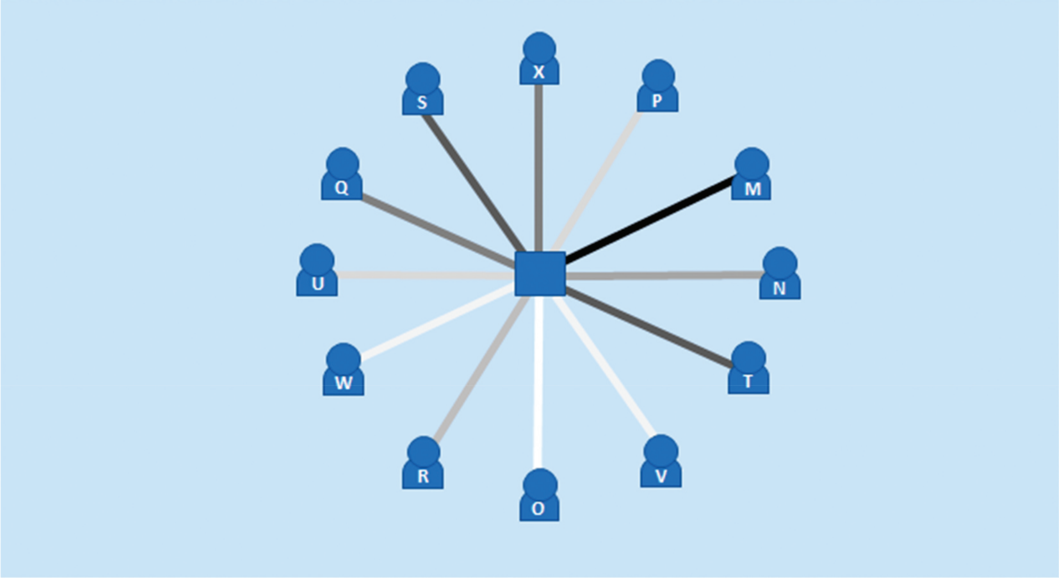

Figure 2a. Network chart for darkness with uncertainty represented as black, white and shades of grey links.

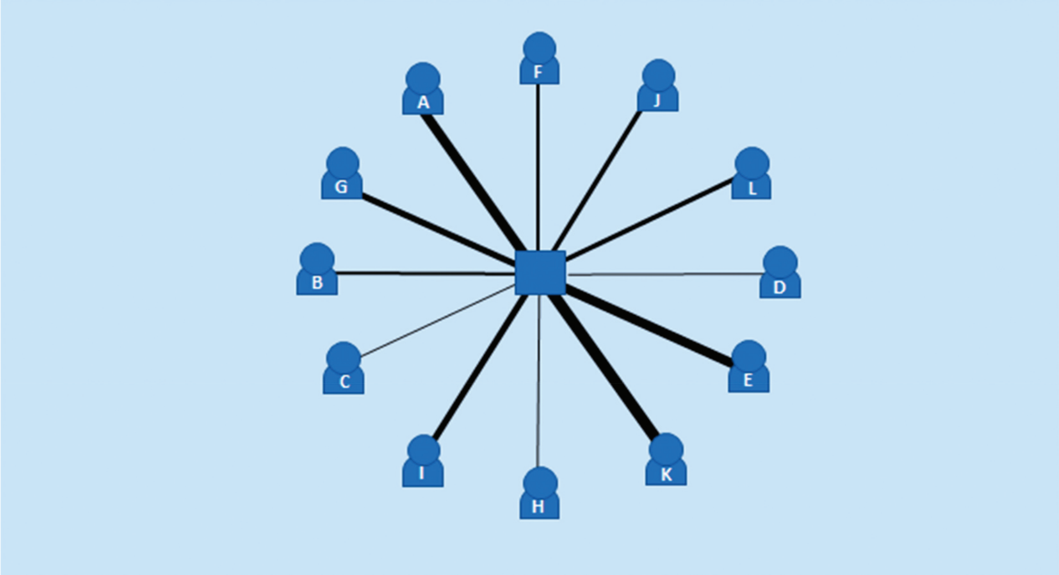

Figure 2b. Network chart for thickness with uncertainty represented as different size links.

information for each link was provided (see Figures 2(a-c). In the thickness condition, the links varied in width (thickness), and in the darkness condition, they ranged from white to black (thickness and darkness were produced via Powerpoint's 'format shape' function, using 'width' and 'colour' options, respectively). In the word condition, the links contained the terms taken from the UK Professional Head of Intelligence Analysis (PHIA)'s 7-category lexicon (also known as the 'yardstick') plus the term 'roughly even chance' taken from the United States Office of the Director of National Intelligence (ODNI) lexicon (see Figure 1). Note that several of the studied terms are common to other lexicons, although their intended (mandated) meaning may differ. This enables us to examine compliance for

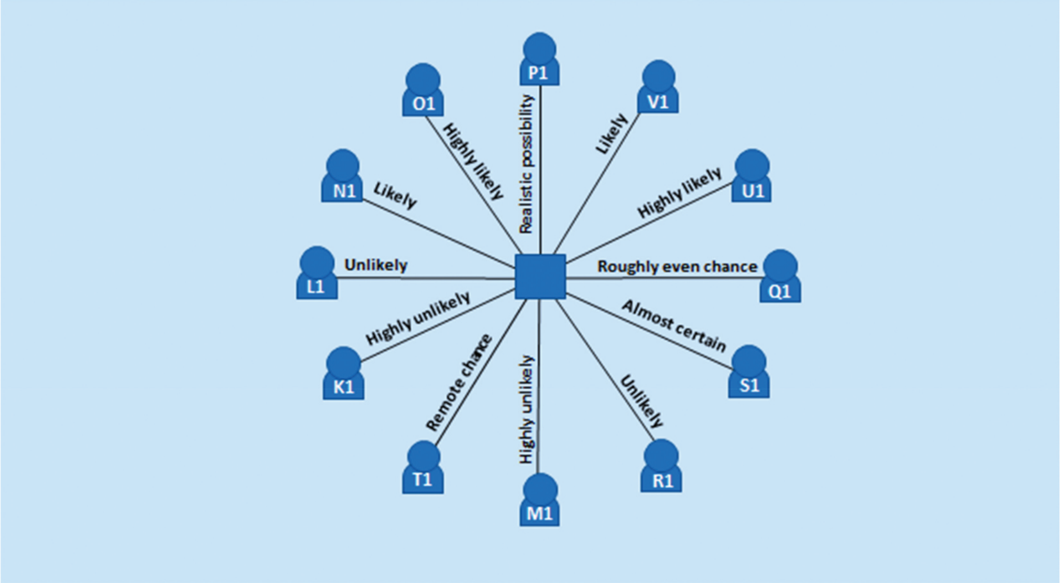

Figure 2c. Network chart for verbal probabilities with uncertainty represented using words.

other lexicons, in addition to PHIA's lexicon, which is the most recently updated,[63] and so arguably the one that could have benefitted most from past research on verbal probabilities.[64]

In order to assess whether visualizations can be intuitively scaled (i.e., if there is agreement across individuals in encoding direction), participants were not instructed on the scale ordering for either of the formats (e.g., they were not told whether darker links represented greater probability). If, for example, most people naturally encode darkness in one direction, then policy-makers could capitalize on this when labelling a visualization accordingly, making the task of interpreting the visual format less cognitively demanding.

Finally, in order to measure test-retest reliability, four of the links were repeated in each task (the same four), with only the letter label being different (e.g., 'A' was re-labelled as 'E'). The four repeats were those lying in-between the end- and mid-points of PHIA's (and others') lexicon i.e., away from the end- and mid-points of the 0–1 probability scale.

Participants were asked to judge the likelihood of each individual being a member of the insurgent group based on the verbal probability describing the link or the darkness or thickness of the link. Judgments were provided on 0–100 per cent scales, numerically marked at 5 per cent-point intervals, and participants were asked to circle one point on each scale.

Procedure

The present study received ethics approval from the first author's University Department Research Ethics Committee. Data was collected by the third author on the first day of training using an individual, self-completion, paper-pencil procedure at the defense training facilities. There was no time limit for completion of the task.

The background information to all tasks stated 'Analytic assessments have a degree of uncertainty associated with them because relevant data may be missing, data collection may be biased, and data may be unreliable as well as purposefully misleading. There are several ways in which this uncertainty can be communicated. Over the page, we present you with three link charts produced by three analysts. The analysts have used different methods for communicating the degree of (un)certainty in their assessments of whether each of 12 individuals are associated with an insurgent group. We want you to tell us how likely you think each individual is to be associated with that group'. Then, each network chart was presented under its respective heading (e.g., 'Chart VS – The uncertainty here is represented using different thickness links', see Figures 2a–2c). All 12 links (including the four repeated ones) were randomly distributed within a chart. The size of the network charts was identical across the three conditions. The order of the conditions was counter-balanced across participants.

Results[65]

The percentage response scale was converted to a 0–1 probability scale for data analyses purposes. Recall that there were eight levels of probability for each format. For the word condition, as in previous research,[66] we used the probability of each level as the mid-point of the numeric range associated with each term in PHIA's lexicon, and .50 for the term *roughly even chance* taken from the ODNI lexicon. Thus, the eight probability levels for the word condition were: .025, .15, .30, .45, .50, .65, .85 and .975. For the visual conditions, each level was coded as a probability ranging from 0 to 1 in even (i.e., .14) increments, which was then rounded to the nearest .05 given that the response scale was numerically marked in such increments. Therefore, the probability levels for the two visual conditions were: 0, .15, .30, .40, .55, .70, .85, 1. All data analyses included only estimates on the first evaluation of items except for the test-retest analysis, and where otherwise indicated. Below, we report how the uncertainty communication formats fared on the various metrics of between- and within-individual performance. Analyses were conducted in R.[67] We used the lmerTest package to run mixed models,[68] and the emmeans package to conduct post-hoc analysis.[69] Three participants did not respond to one or more links (for an unknown reason), and so their data are missing on these trials.

Signal detection theory emphasizes sensitivity (also known as information acquisition). One way to measure sensitivity to the information is the use of a measure called d'. Another approach is to measure slopes from a regression model, which are equivalent to d'.[70] Below, we calculated slopes from linear regression models for each participant, for each condition to firstly examine the agreement in encoding direction, and to secondly measure sensitivity to the probability being communicated. In these regression models, the dependent variable was estimated probability, and the independent variable was probability level. We ran separate models for each participant, for each condition.[71]

Agreement in encoding direction

From the regression models, we assessed the direction of the slopes. For the word condition, positive slopes indicate interpreting terms such as *likely* as being more probable than terms such as *unlikely*.[72] As the lines in Figure 3 show, all but one participant (98 per cent) had a positive slope in the word condition.

In the two visual conditions, recall that one aim was to examine how individuals (naturally) preferred to interpret or encode increments in the darkness and thickness of lines. Thus, participants were not provided any explicit instructions about whether darker or thicker links referred to higher versus lower probability levels. In other words, participants were not told how to encode the darkness and thickness of links, so either mapping is technically correct within our task. We therefore assessed participants' natural inclinations in how they interpreted line darkness and thickness and called the measure *agreement in encoding direction*. As shown in Figure 4, there was a clear preference to encode darker links (81 per cent of participants) and thicker links (93 per cent of participants) as indicating greater probability (i.e., more participants had positive than negative slopes as shown in Figure 3).

All conditions were significantly different from each other in terms of the number of participants who had positive versus negative slopes.[73] Note that in further analyses we standardized any differences across participants in encoding direction by assessing absolute slopes for the visual conditions because there is ambiguity in how darker or thicker lines should be interpreted. Given that this is not the case for the word condition, we used signed slopes for this condition.

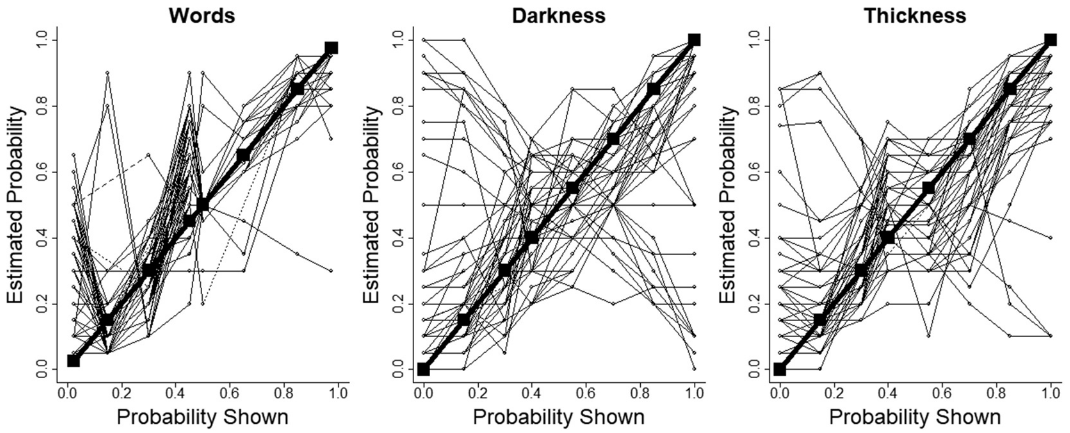

Figure 3. Estimated probability as a function of probability being communicated for each condition. Each line = one individual. Dotted lines were plotted when there was a missing value. Thick lines = the accurate response as specified by the probability levels defined above.

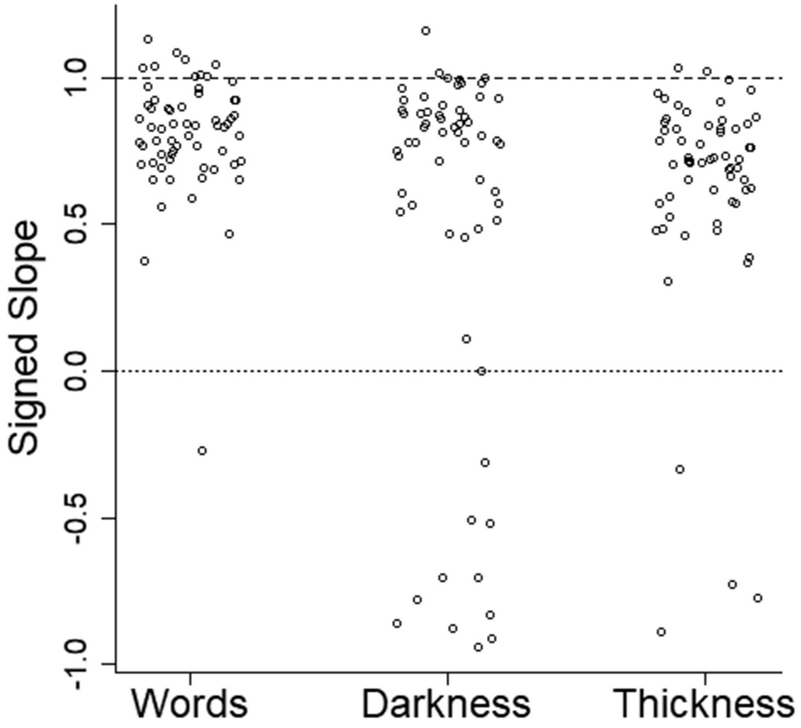

Figure 4. Signed slopes for each condition. Each dot = one individual. A slope of one (dashed line) indicates perfect sensitivity and a score of zero (dotted line) indicates chance performance (sensitivity scores were based on signed slopes for the word condition and absolute slopes for the visual conditions; see texts for details). Shorter solid lines indicate the mean for each condition. Some jitter was added to the abscissa to improve visibility.[74].

Sensitivity to probability being communicated

A primary goal of probability communication is to convey a probability value unambiguously so that it is interpreted as intended. According to slope analysis, such sensitivity can be measured as the magnitude of the slope. We refer to this measure as a *sensitivity score*, where a score of 1 indicates perfect sensitivity and zero indicates chance performance. For illustration, the thick lines and large symbols in Figure 3 show ideal performance and correspond to a sensitivity score of 1. Since no directions were given to participants regarding how thicker or darker links should be interpreted, the sensitivity scores for these conditions were calculated as absolute slopes. In the word condition, there was no such ambiguity as to how words should be interpreted, so we used the signed slopes from this condition. We analysed sensitivity scores (which, as we mentioned earlier, were absolute slopes for the visual conditions and signed slopes for the word condition).[75] The mean sensitivity scores across participants for each condition were as follows: for the word condition ($M = 0.80$, $SD = 0.20$), for the darkness condition ($M = 0.77$, $SD = 0.21$) and for the thickness condition ($M = 0.72$, $SD = 0.17$). In further examination of means, we found that sensitivity was significantly greater under the word than thickness condition, but not significantly greater under the word than darkness condition, and that sensitivity was not significantly different between the darkness than thickness conditions.[76] For all three conditions, sensitivity was significantly less than the optimal score of 1 as revealed by separate one-sample *t*-tests for each condition, *p*s < .001, *d*s >3.6.[77]

Inconsistency in visual scale usage

People can be good at estimating the lowest probability as low and the highest probability as high. These estimates would lead to high sensitivity scores even if the person did not have good sensitivity to probabilities in-between the extremes or end-points of the probability scale. Given that many probabilities are likely to occur in the middle, instead of at the extremes, it is important to assess which format is most effective at conveying these intermediary probabilities. One way to do this is to compare the verbal and visual formats in terms of whether participants responded to each using equally-sized probability intervals i.e., whether estimations increased consistently as the probability being communicated increased.

Some participants increased their estimates in nearly even increments whereas some showed inconsistent increases, with some patterns of responses resembling a staircase function (notice the horizontal lines in the Figure 3 panel for the thickness condition between the middle probability levels of .4 and .55). A staircase function implies that individuals could not differentiate the probabilities between those two corresponding probability levels for which the horizontal part of the staircase occurred. In the case of the example of the middle probability levels for the thickness condition, each horizontal line indicates an instance for which a participant made the same estimates when the probability level was .4 and .55. We quantified this behavior as a measure we call *inconsistency in scale usage*.

Inconsistency in scale usage was assessed as follows: we calculated difference scores as the difference in estimated probability at each successive interval of the probability being communicated. This was done for each participant for each visual condition because these two conditions had equally-sized incremental increases in probability. We then calculated the standard deviation (*SD*) of these difference scores for each participant for each visual condition. Greater inconsistent estimates across intervals would lead to higher *SD*s whereas less inconsistent estimates would lead to lower *SD*s.[78]

We found a significant difference across the two visual conditions in terms of *inconsistency in scale usage* (see Figure 5).[79] Participants were 54 per cent more inconsistent in their estimates in the thickness condition than in the darkness condition i.e., they were more likely to increase their estimates in uneven intervals (inconsistent) as the probability being communicated increased in the thickness condition.[80] In other words, this suggests better consistency for the darkness condition than the thickness condition.

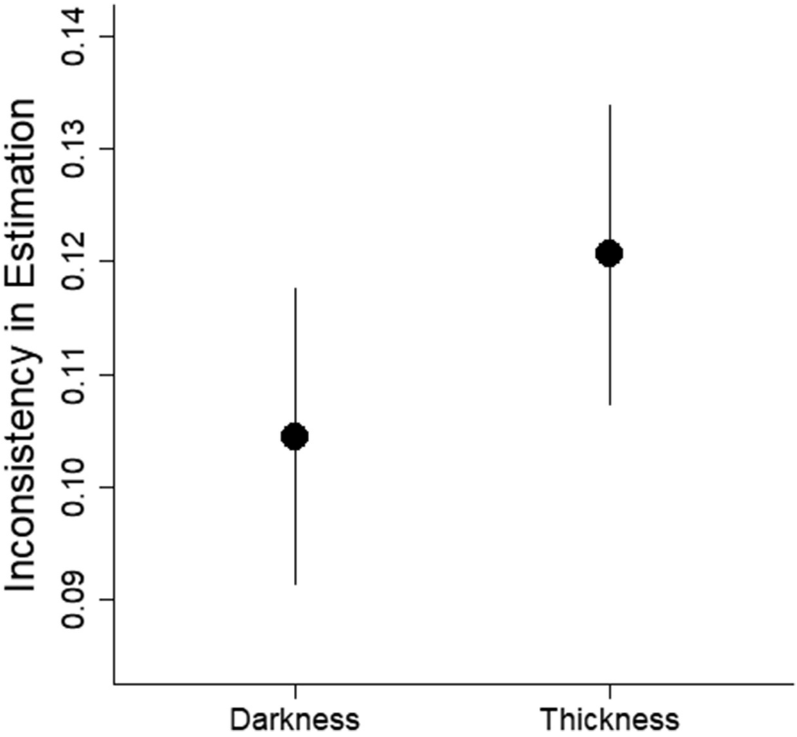

Figure 5. Inconsistency in scale usage across probability intervals for each visual condition. Inconsistency was measured as the standard deviation (*SD*) of differences in estimated probability across successive increments of the probability being communicated. Higher values indicate greater inconsistency (i.e., worse performance). Error bars represent 95 per cent CIs.

Inter-individual variability in estimation

Inter-individual variability in estimation for each condition was measured in terms of the range of probability estimates provided across participants at each probability level (i.e., highest estimate minus lowest). The range could vary from 0 to 1. A higher range signifies greater variability across participants. For example, in the word condition in Figure 2, the first term (*remote chance* which according to PHIA is intended to represent a mid-point of .025 (along a .01 to .05 interval), is interpreted by participants as being anywhere from .05 to .65, indicating a range of .60.

Figure 6 shows the range of estimates provided at each of the eight probability levels for each condition. Overall, the mean range across all probability levels for each condition was as follows: word condition (*M* = .64, *SE* = 0.06), darkness condition (*M* = .56, *SE* = 0.06) and thickness condition (*M* = .53, *SD* = 0.06). Further analyses showed that the differences in range across the three conditions were not significant, and pairwise comparisons between the conditions similarly showed no significant differences.[81]

Intra-individual variability in estimation

Intra-individual variability in estimation was measured using *test-retest reliability*. Recall that for each condition, four of the links in the network chart were repeat presentations (unbeknownst to participants). On average, across participants, on the second evaluation of an item, the same estimate was provided 84 per cent of the time in the word condition, 63 per cent of the time in the darkness condition, and 54 per cent of the time in the thickness condition. The difference

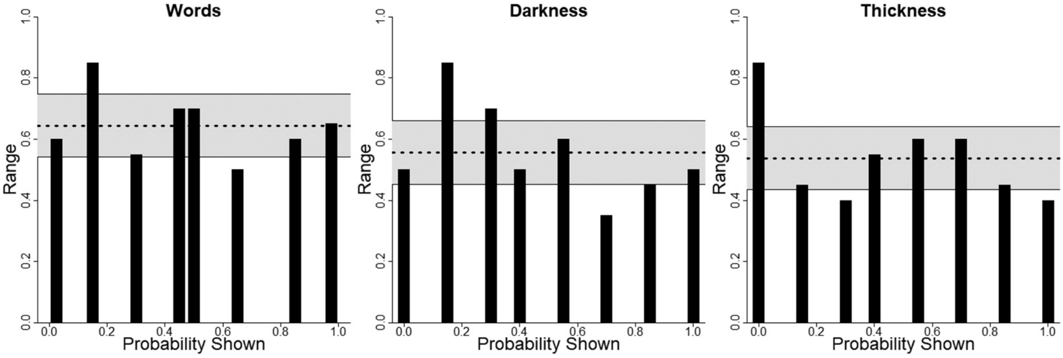

Figure 6. The range of estimated probability at each level of probability being communicated for each condition. Dashed horizontal lines = mean range for each condition, and shading = 95 per cent CIs.

between the word condition and the thickness condition was significant. The difference between the word and darkness conditions was not significant, and neither was the difference between the darkness and thickness conditions.[82]

More specifically, Figure 7 shows each participants' difference scores (i.e., the first estimate minus the second) for each of the four repeated items, for each condition. The mean absolute difference across participants and repeated items for the word condition was 0.03 (95 per cent CI [.02, .04]). For the darkness condition, it was 0.04 (95 per cent CI [.03, .05]), and for the thickness condition it was 0.05 (95 per cent CI [.04, .06]). The absolute difference for the word condition was significantly smaller than for the thickness condition, but not significantly different from the darkness condition. The absolute difference between the darkness and thickness conditions was not significant.[83]

Further analysis

To further explore whether some practice with the visual formats led to rapid increases in performance, sensitivity scores were calculated for the repeated items. Sensitivity scores (calculated as slopes from linear regressions for which signed slopes were used for the word condition and

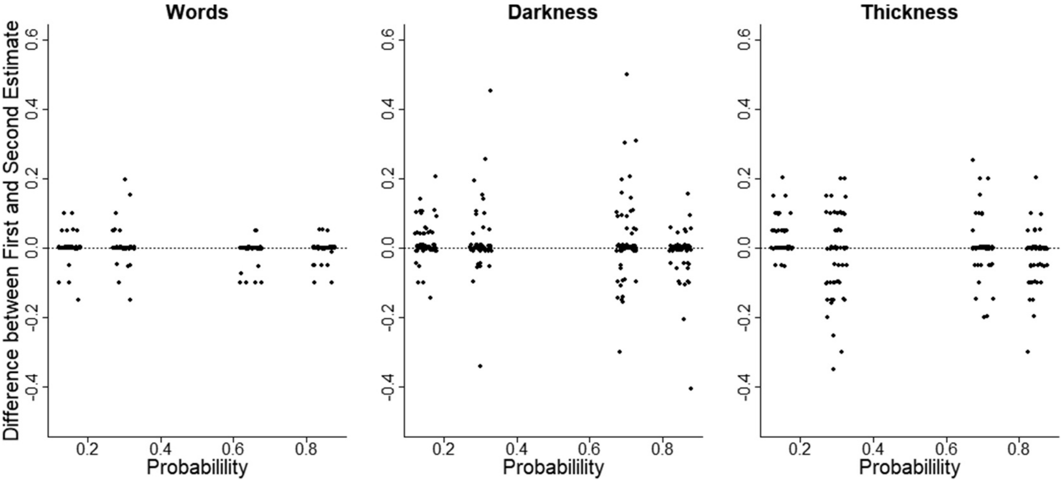

Figure 7. First estimate minus second for each probability being communicated for the four repeated items in each condition. Each dot = one individual. Dashed horizontal line at 0 = no difference between estimates. Some jitter was added along the abscissa to improve visibility.

absolute slopes were used for the visual conditions, as before) were calculated for the first evaluation and for the second evaluation for each participant, for each condition. For both visual conditions, mean sensitivity scores across participants were better for the second evaluation compared with the first (darkness: .81 vs .82, and thickness: .83 vs .85; first evaluation then second evaluation respectively), although the difference did not reach statistical significance for either condition.[84] For the word condition, sensitivity scores were worse for the second evaluation (1.03 vs 1.05; first evaluation then second evaluation respectively), but again not significant.[85] As a reminder, sensitivity scores can be greater than 1 when middle-to-low probabilities are underestimated and middle-to-high probabilities are overestimated.

Compliance with intelligence community lexicons

Finally, data from the word condition was used to examine participants' compliance with the mandated meaning of terms in various intelligence community lexicons. As mentioned earlier, some of the terms in PHIA's lexicon are also present in other lexicons currently in use by Western intelligence communities, although they may be intended to convey different numeric probability values (see Figure 1). Figure 8 shows participants' estimates (dots) of each term, and for illustration we present the intended range of the terms as specified by PHIA, along with the range for *roughly even chance* specified by ODNI. Note, that interested readers can also make a comparison with the intended ranges for the other terms in the ODNI lexicon as well as for those terms that are in the NATO lexicon (see Figure 1).

Overall, it can be seen that at the lowest and highest probability levels, participants' estimates were characterized by great response compression (i.e., overestimation at the lowest level and underestimation at the highest level). In addition, whereas most participants' estimates of the mid-lexicon term in the ODNI lexicon (i.e., *roughly even chance*) and the NATO lexicon (i.e., *even chance*) was as intended at .50, participants greatly overestimated the probability for *realistic possibility*, which is the mid-lexicon term used by PHIA.

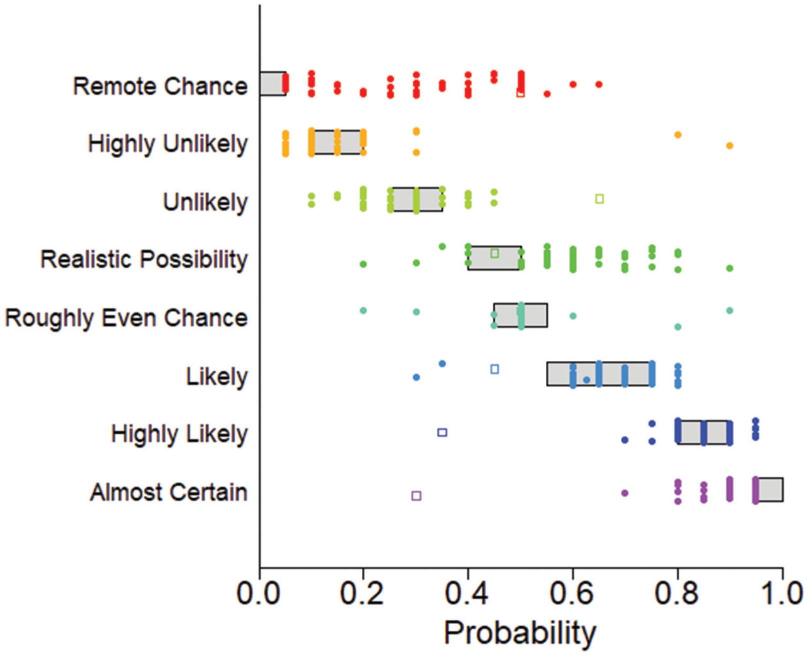

Figure 8. Individuals' (dots) estimates of each word plotted in relation to the range of probabilities assigned to each word in the PHIA lexicon and 'roughly even chance' taken from the ODNI lexicon (boxes).

Table 1. Median interpretations of words across individuals and mid-point intended meanings of the words according to current intelligence community lexicons.

	Dutch sample	PHIA (UK)	ODNI (US)	NATO
Remote chance	.30	.025	.025	
Highly unlikely	.10	.15		.05
Unlikely	.30	.30	.325	.25
Realistic possibility	.60	.45		
Even chance	.50		.50	.50
Likely	.70	.65	.675	.75
Highly likely	.85	.85		.95
Almost certain	.95	.975	.955	

To further explore the extent to which participants interpreted the terms studied here as intended by the various lexicons, Table 1 presents the median interpretations of the terms across participants alongside the mid-point of the range of probabilities they are intended to represent in the various lexicons. The over-estimation at the lowest probability levels is evident for all three lexicons, and the under-estimation at the highest probability level is evident for the NATO lexicon. Again, there is greater compliance with the mid-lexicon term used by NATO and ODNI, while there is over-estimation of the mid-lexicon term used by PHIA.

Discussion

Even well-informed and reasoned analytic judgments can lead to poor outcomes if the probabilities (or uncertainty) associated with them are misunderstood by decision-makers. Despite psychological research pointing to the shortcomings of using verbal probabilities to communicate uncertainty unambiguously,[86] Western defense and security organizations rely on this format.[87] Perhaps unsurprisingly therefore, recent research has demonstrated the ineffectiveness of various lexicons used by the intelligence community.[88]

The main goal of the present study was to explore the value of a potential alternative approach by comparing the use of verbal probabilities against visual encoding channel-based representations of uncertainty (i.e., darkness and thickness). This bypasses the apparent reluctance that the intelligence community has with using numeric probabilities[89] and builds on the common use of visualizations in technology-aided defense and security analytics.[90] In addition, although others have previously proposed the use of visualizations to communicate uncertainty in intelligence analysis[91] there has been no systematic examination of the use of verbal probabilities versus visualizations in this domain. Our research represents one exploratory step in this direction.

We appreciate that any efforts to empirically explore alternatives to current policy and practice in an expert/professional domain such as intelligence analysis will necessarily imply asking participants to respond to stimuli that they are familiar with (i.e., existing policy) and stimuli that are novel to them (i.e., potential policy alternatives). To overcome this problem, researchers could employ non-expert/lay samples who would be unfamiliar with all such stimuli, however, this comes at the charge of being unable to generalize to the relevant participant population. The present research therefore, involved a professional sample of analysts, given that they are the typical consumers of analytic products, along with policy- or decision-makers (whom we did not have access to). The communication formats (i.e., verbal versus visual representations i.e., darkness and thickness) were assessed on various metrics of between- and within-individual performance. Below, we discuss the findings as well as their implications and provide suggestions for future research on this topic.

Visualizing versus verbalizing uncertainty

The word format led to greater agreement as to which words signaled greater probability (e.g., 'likely') compared to whether darker or thicker links signaled greater probability. This is unsurprising because darkness and thickness have no intrinsic meaning and so are not explicitly scaled, making them difficult to evaluate, according to the 'evaluability hypothesis'.[92] In order to assess whether the two visual formats can be intuitively scaled, participants were not instructed on the scale ordering for darkness and thickness of links. If most people naturally encode a visualization in one direction then policy-makers could capitalize on this, and label it accordingly, thus making the task of interpreting the visual format less cognitively demanding. We found that the majority of individuals agreed that darker and thicker links represented greater probability. These findings contribute to past research involving simple comparison techniques to investigate darkness and thickness,[93] and studies showing that darkness is intuitively associated with greater probability.[94]

We also found that in terms of sensitivity to the probability being communicated, performance under the word condition was significantly better than the thickness condition, but not significantly different to the darkness condition. Performance under the two visual conditions was not significantly different. Overall, however, for all three conditions sensitivity was less than the optimal score of 1. This finding underscores the challenges associated with communicating uncertainty.

Variability in estimation of the probability being communicated either at the *inter*- or *intra*-individual level poses serious problems for the effectiveness of an uncertainty communication policy. In the present study, both the verbal and visual representations of uncertainty demonstrated considerable *inter*-individual variability as measured by the range of probability estimates provided across individuals, and no one format outperformed the others. These findings are compatible with past research demonstrating inter-individual variability in interpretation of verbal probabilities,[95] including terms in the lexicons used by the intelligence community.[96] Similarly, the findings also point to the difficulty that visual encoding channel-based representations (i.e., darkness and thickness) may have in communicating uncertainty across people, an issue that Hullman et al. state has garnered little research attention to-date.[97]

Another issue that has not been sufficiently addressed by researchers examining uncertainty visualization is *intra*-individual variability. We measured test-retest reliability and found it to be significantly greater for the word condition than for the thickness condition, but not significantly different from the darkness condition. One potential explanation for part of these findings is that individuals generally have more practice mapping words to numbers than mapping visualizations (particularly thickness) to numbers. More research is needed to understand the effect of exposure driven sensitivity and the maximum level of performance that can be achieved with the visual format. Indeed, according to Hullman et al'.s review of uncertainty visualization evaluation studies published since 1993, the issue of learnability has yet to be empirically addressed, even though such research would be useful for organizations deciding amongst uncertainty communication policies.[98]

For now, returning to the test-retest performance for words, it is worth noting that this was less than perfect (i.e., 84 per cent). The words that were repeated referred to terms that lay in-between the mid- and end-points of the probability scale (i.e., *highly unlikely*, *unlikely*, *likely*, *highly likely*), and as other research has found, these intermediate points are vaguer in peoples' minds.[99]

Darkness versus thickness

When comparing the two visual formats, darkness performed marginally better than thickness in terms of sensitivity to the probability being communicated, although this difference did not reach statistical significance. Two different thicknesses were somewhat more likely to be estimated as being equivalent (in probability), so increments in thickness were somewhat less easily differentiated than increments in darkness. Whereas line colour has natural end points (black and white), thus making levels of colour easier to evaluate, especially when presented simultaneously (or jointly) as in the present study, line thickness does not. Past research has documented the superiority of darkness over other visual formats,[100] including thickness,[101] and the evaluability of darkness may be one potential explanation. We examined thickness in terms of width, and so future research could explore whether other operationalizations such as coverage, which has natural end points (all or none) and so easier to evaluate, might fare better. Drecki, for instance, found that coverage was more effective in communicating uncertainty than height, which, like thickness, does not have natural end points.[102] Finally, it is worth noting that we do not anticipate that using darkness and thickness at the same time would necessarily be better than either one on its own. This is because the perception of colour is known to be affected by size.[103]

Compliance with intelligence community verbal probability lexicons

A secondary aim of the present study was to examine participants' compliance with existing standardized lexicons (i.e., the extent to which users interpreted terms as mandated). Whereas past research on this issue has been conducted in native English-language speaking countries,[104] we focused on a country that has its own language (Dutch). We observed that Dutch analysts' and officer cadets' estimations of the English-language verbal probability terms were characterized by overestimation at the lowest probability level and underestimation at the highest level, such that their interpretations fell outside of the ranges intended by the various lexicons used by the intelligence community (see Figure 1). This regression to the mid-point of the scale has previously been observed in English-language speaking analysts.[105] It has also been shown among non-native language speakers in other domains where English terms were translated into participants' native language,[106] pointing to the ubiquity of this 'regression' problem.

In particular, our analysis revealed that although *remote chance* is used to represent the smallest probability in both PHIA's and ODNI's lexicons, participants, on average, interpreted the term as .28, which is a great overestimation of the maximal value of .05 in these lexicons. Ho et al. similarly reported an overestimation of this term among UK and Canadian analysts.[107] In addition, we found that the average estimation of NATO's lowest ranked term (i.e., *highly unlikely*) was greater than intended. Together, these findings suggest that none of the lexicons currently in operation in the intelligence community are capable of describing very low probabilities.

Beyond, misinterpretation at the end points of the probability scale, we also observed that whereas participants' estimates of terms fell within the very broad category ranges used by NATO, their estimates fell outside the narrower ranges used by PHIA and ODNI. Of particular concern is the term *realistic possibility* which, according to PHIA, should be used to communicate .40 to .50, but was greatly overestimated in the present study, with the average interpretation being .60. Others have expressed concern with the use of this term, and Barnes 'banned' its use in his Canadian military intelligence unit.[108] In a study of UK analysts, Dhami reported that *realistic possibility* did not appear in the lexicons of any of her sample, and so one explanation for the present findings may be that individuals were unfamiliar with the term.[109] Another explanation lies in the observation that *possible* is very broadly interpreted.[110] *Possible* was used to represent from 10 per cent to 60 per cent in Dhami's study, and when this term is translated into Dutch, Willems et al. found that it represented from 20 per cent to 70 per cent.[111] In addition, Renooij and Wittman reported that the Dutch equivalent of *possible* (i.e., mogelijk) was assigned a value of .86, which falls outside the range in Willems' et al. study.[112] Together, these findings suggest that PHIA ought to refrain from using the term *realistic possibility*.

It is worth pointing out that, in practice, users of analytic products may have a 'guideline table' available to them which contains numeric translations of terms, and our study did not contain such a table in the word condition. It is unknown to what extent such a table would have improved compliance with the intelligence community lexicons studied here. We refrained from providing such a table because it would mean we could only comment on one lexicon that is currently in use rather than the three we have presently evaluated as three different tables would have been needed, and it would have limited our ability to directly compare our results with those of past research involving analysts which mostly does not provide such tables. Efforts to improve compliance with lexicons used in other domains, including providing numeric translations in brackets alongside the terms in text,[113] making a 'guideline table' available, or introducing a 'tooltip' where numeric translations appear on screen when readers hover over a term, suggests that the improvement in compliance is moderate and that not all users engage with such interventions.

In the present study, non-native English-speaking participants provided probability estimates on the basis of verbal probabilities presented in the English language. The findings confirmed those of past research examining native English-language speaking analysts' compliance with standardized lexicons,[114] and contribute new insights into the difficulties that non-native English-language speaking allies may have when consuming intelligence products using terms in these lexicons. Other research suggests that the problems associated with the terms studied here are not overcome by simply translating them. For instance, according to Willems et al., when *unlikely*, *likely* and *almost certain* are translated into Dutch, the average interpretation across a sample of nearly 900 people remains either much lower (i.e., 16 per cent for *unlikely* and 88 per cent *almost certain*) or higher (i.e., 75 per cent for *likely*) than typically intended by the various lexicons used in intelligence domain.[115] Therefore, an approach that does not rely solely on simple translation is required (including the efforts involving numeric translations mentioned above). Ho et al, for instance, propose developing probability communication lexicons based on empirical evidence showing how people numerically interpret specific phrases, and selecting phrases with little or non-overlapping interpretations.[116] They have demonstrated the efficacy of their approach in an intelligence analysis context (Study 2).

Avenues for future research

The present study responds to calls to employ an evidence-based approach to policy and practice in intelligence analysis[117] as well as to calls for more rigorous empirical tests of visual representations of uncertainty.[118] Future research could compare verbal probabilities and darkness and thickness against other metrics. For instance, participants in the present study were asked to provide point estimates rather than intervals and although past research demonstrates the vagueness of probability terms in the minds of individuals,[119] it is unknown how this compares with the imprecision associated with interpreting darkness and thickness. We found that people may have problems with perceiving different thicknesses and so thickness may not perform as well as other representations in terms of imprecision, which is an important metric, given that imprecision can increase opportunities for miscommunication and misunderstanding. In addition, obtaining interval estimates enables measurement of 'agreement' with intended numeric ranges associated with words.[120] Another metric that could be used in future research is how well the verbal and visual formats fare in terms of the effect on decision-making. Indeed, by contrast to the amount of literature examining how people interpret verbal probabilities, there is relatively less work on the effects of these interpretations on decisions,[121] especially compared to visual representations of uncertainty.

Future research should also compare the verbal format against other visual representations such as fuzziness, transparency and location,[122] as well as when both formats are presented in a separate evaluation mode. Not only would this open up the possibility for using a greater range of visualizations to communicate uncertainty in intelligence analysis, but this would also help to more confidently draw conclusions about the relative efficacy of the verbal and visual formats, and the best approach for their presentation.

Researchers may wish to explore the efficacy of a combined approach, examining whether adding visual cues to probability terms can increase interpretation accuracy and consistency, although Edwards and Nelson did not find such a combination to be particularly helpful.[123] In addition, future research could investigate the extent to which interpretations under the different formats are affected by the context in which they are presented. Past research has revealed context effects on the interpretation of verbal probabilities,[124] including in the intelligence analysis domain,[125] and the extent to which these effects also occur for visual representations remains to be known.

Finally, although some in the intelligence community may rail against the use of numbers,[126] it is important to know the extent to which both verbal and visual formats fare against the numeric format. This is an issue that has yet to be empirically examined as prior research has typically compared either the verbal and numeric format[127] or the visual and numeric format.[128] Answers to these questions can be used to inform the development of more effective uncertainty communication policies in the intelligence analysis domain, where the simple miscommunication of uncertainty can result in deleterious consequences.

For now, leveraging the insights drawn from the present research for intelligence practices can take various forms. For instance, the traditional textual format of intelligence assessments, such as the National Intelligence Estimates key judgements in the United States, could be adapted by highlighting words of estimative probability through variations in thickness and darkness. The uptake of such an intervention, as well as any others would need to be informed by further research as suggested, as well as the need to consider whether the benefits of more unambiguous uncertainty communication outweigh the practicalities of their implementation. We believe they do.

Notes

1. e.g. Budescu, Por & Broomell, “Effective Communication”; Dhami & Mandel, “Words or Numbers?”; and Hart et al. “Guidance on Communication”.
2. We use the terms probability and uncertainty interchangeably as in much of the literature on uncertainty communication generally (e.g., see Collins & Hahn, “Cultivating Credibility with Probability Words and Numbers”) and the literature on communicating uncertainty in intelligence analysis more specifically (for a review see Dhami & Mandel, “Words or numbers?”).
3. e.g. Butler, *A Review;* Chilcot, *The Report*; United States Congressional Select Committee on Intelligence. *U.S. Intelligence Community’s Prewar Assessments*; and United States Central Intelligence Agency, *Report on Intelligence Judgements.*
4. For a review see Dhami & Mandel, “Words or Numbers?”
5. ibid.
6. NATO Standardization Office. *AJP-2.1*; and US ODNI *Prospects for Iraq’s Stability*; College of Policing. *Delivering Effective Analysis.*
7. For a review see Dhami & Mandel, “Communicating Uncertainty.”
8. See Dhami & Mandel, “Words or Numbers?”
9. Lonsdale & dos Santos Lonsdale “Handling and Communicating.”
10. Singh & Philip, “An Innovative Scheme.”
11. Gentry & Gordon, *Strategic Warning Intelligence*.
12. Chung & Wark, “Visualising Uncertainty.”
13. E.g. Weiss, “Communicating Uncertainty”.
14. College of Policing. *Delivering Effective Analysis.*
15. See Dhami et al “Improving Intelligence Analysis.”
16. E.g. Wallsten et al.,“Preferences and Reasons”; and Juanchich & Sirota, “Do People Really Prefer.”
17. Barnes, “Making Intelligence Analysis”; Friedman & Zeckhauser, “Assessing Uncertainty”; and Marchio, “If the Weatherman Can.”
18. Zimmer, “Verbal vs. Numerical”; and Zimmer, “A Model.”
19. e.g. Beyth-Marom, “How Probable is Probable?”; Budescu et al., “The Interpretation”; Clarke et al., “Ratings of Verbal Expressions”; Dhami, & Wallsten, “Interpersonal Comparison”; Lichtenstein & Newman, “Empirical Scaling”; and Wiles, Duffy & Neill, “The Numerical Translation.”
20. Dhami, “Towards an Evidence-Based Approach”; Ho et al., “Improving the Communication”; Johnson, *Numerical Encoding*; Mandel, “Accuracy of Intelligence Forecasts”; Mandel & Irwin, “On Measuring Agreement”; Mandel & Irwin, “Tracking accuracy”; Wallsten et al., “Exploring intelligence Analysts’ Selection”; Wark, “The Definition”; and Wintle et al., “Verbal Probabilities”.
21. Wark, “The Definition.”
22. Budescu et al., “The Interpretation”; Mandel & Irwin, “On Measuring Agreement”; and Wintle et al., “Verbal Probabilities.”

23. e.g. Budescu, Por & Broomell, "Effective Communication"; Budescu et al., "The Interpretation"; Janzwood, "Confident, Likely, or Both?"; Mandel & Irwin, "On Measuring Agreement"; Milne et al., "Communicating the Uncertainty"; Wallsten, Fillenbaum & Cox, "Base-Rate Effects"; and Wintle et al., "Verbal Probabilities."
24. Kesselman, *Verbal Probability Expressions*.
25. Dhami, "Towards an Evidence-Based Approach"; Ho et al., "Improving the Communication"; Mandel, "Accuracy of Intelligence Forecasts"; Mandel & Irwin, "On Measuring Agreement"; and Wintle et al., "Verbal Probabilities."
26. e.g. Budescu et al., "The Interpretation"; Harris et al., "Lost in Translation?"; and Kolesnik, Silska-Gembkaa, & Gierusz, "The Interpretation."
27. Ho et al., "Improving the Communication."
28. Mandel et al., "Arithmetic Computation"; and Teigen, Juanchich & Løhre, "Combining Verbal Forecasts."
29. Mandel et al., "Arithmetic Computation."
30. Collins & Mandel, "Cultivating Credibility"; see also Teigen & Brun, "Yes, but it is Uncertain"; and Teigen & Brun, "The Directionality."
31. e.g. Barnes, "Making Intelligence Analysis"; Dhami & Mandel, "Words or Numbers?"; Friedman et al., "The Value of Precision"; Johnson, *Numerical Encoding*; and Marchio, "If the Weatherman Can."
32. Dhami & Mandel, "Words or Numbers?"
33. US ODNI *Prospects for Iraq's Stability*.
34. Kent, *Words of Estimative Probability*.
35. Reyna & Brainerd, "Numeracy, Ratio Bias."
36. e.g. Weiss, "Communicating Uncertainty."
37. College of Policing. *Delivering Effective Analysis*.
38. Padilla, Kay & Hullman, *Uncertainty Visualization*.
39. e.g. Deitrick & Wentz, "Developing Implicit Uncertainty."
40. A hybrid approach is also possible (e.g., probability density and interval plot). See Fernandes et al., "Uncertainty Displays"; and Padilla, Kay & Hullman, *Uncertainty Visualization*.
41. e.g. Kinkeldey, MacEachren & Schiewe "How to Assess"; and MacEachren et al., "Visual Semiotics."
42. See Spiegelhalter, Pearson & Short, "Visualizing Uncertainty."
43. e.g., for a review see Garcia-Retamero, Okan & Cokely, "Using Visual Aids."
44. e.g. Kinkeldey et al. "Evaluating the Effect."
45. e.g. Lavine & Gouin, *Applicability*; McCue, "Security Threats"; Phillips et al., "Applying Social Network Analysis."
46. Larkin & Simon, "Why a Diagram."
47. e.g. McKenzie et al., "Assessing the Effectiveness."
48. e.g. Dong et al., "Uncertainty Visualization."
49. Turkay et al., "Special Issue."
50. Correll, Moritz, & Heer, "Value-Suppressing Uncertainty Palettes."
51. Hsee, "The Evaluability Hypothesis"; Hsee et al., "Preference Reversals"; Hsee & Zhang, "General Evaluability Theory"; and Dhami, & Wallsten, "Interpersonal Comparison."
52. Bisantz et al.,"Visual Representations"; Boukhelifa et al. "Evaluating Sketchiness"; Holliman et al., *Visual Entropy*; and Kunze et al., "Augmented Reality Displays."
53. e.g. Cheong et al., "Evaluating the Impact"; Korporaal, Ruginski & Fabrikant, "Effects of Uncertainty Visualization"; and Taylor, Dessai & Bruine de Bruin, "Communicating Uncertainty."
54. e.g. Boukhelifa et al. "Evaluating Sketchiness."
55. Hogan Carr et al., "Effectively Communicating."
56. Milne et al., "Communicating the Uncertainty."
57. Hullman et al., "In Pursuit of Error"; Jena et al.,"Uncertainty Visualisation"; and Kinkeldey, MacEachren & Schiewe "How to Assess."
58. e.g. MacEachren et al., "Visual Semiotics."
59. e.g. Drecki, "Visualization of Uncertainty"; Leitner & Buttenfield, "Guidelines"; and Sanyal et al., "A User Study."
60. Dhami, "Towards an Evidence-Based Approach"; Ho et al., "Improving the Communication"; and Mandel, "Accuracy of Intelligence Forecasts."
61. Renooij & Witteman, "Talking Probabilities."
62. This was one of a set of six separate studies that the sample participated in. The studies were divided into two separate blocks and this study was the last of block one, with the preceding study being on an unrelated topic.
63. Marchio, "If the Weatherman Can. . ."
64. Dhami & Mandel, "Words or Numbers?"
65. The data and the analysis scripts can be found at https://osf.io/pku4n/.
66. e.g. Mandel & Irwin, "Facilitating Sender-Receiver Agreement"; and Mandel & Irwin, "On Measuring Agreement."
67. R Core Team, "R".
68. Kuznetsova, Brockhoff & Christensen, "lmerTest Package."
69. Lenth, "Emmeans".

70. DeCarlo, "Signal Detection Theory."
71. Much of the analyses involved linear mixed models. Running power analyses on linear mixed models requires estimates of variance that were not yet known due to the novelty of our research aims and analyses. Therefore, instead we ran a power analysis as if we had run paired sample *t*-tests. Given our sample size of 62, we had 80 per cent power to detect an effect size of Cohen's $d = .36$, which is a small-to-medium effect size. To detect a medium effect ($d = .5$), we had 97 per cent power, and to detect a small effect ($d = .2$), we had 34 per cent power.
72. To recall, estimated probability was the dependent measure, and probability level was the independent measure.
73. A general logistic mixed model was used to make this assessment where the dependent measure was slope direction (coded as 1 for positive and 0 for negative), the fixed effect was condition, and the random effect was participant. Comparisons showed significant differences across all conditions, all *p*s < .001. The comparison between the word condition and darkness conditions was $z = 7.73$, $p < .001$. The comparison between the word and thickness conditions was $z = 6.50$, $p < .001$. The comparison between the darkness and thickness conditions was $z = -4.70$, $p < .001$.
74. Some individual sensitivity scores were greater than 1. Given that a score of 1 is ideal performance, it may seem strange to observe scores greater than 1. This occurs when lower probabilities are underestimated, and higher probabilities are overestimated.
75. A linear mixed model was computed with sensitivity score as the dependent measure, condition as the fixed effect, and participant as the random effect.
76. For the word compared to thickness condition, $t = 3.09$, $df = 121$, $p < .01$, estimate = 0.09, $SE = 0.03$, $d = .44$. For the word compared to the darkness condition, $t = 1.24$ $df = 121$, $p = .22$, estimate = 0.035, $SE = 0.03$, $d = .18$. For the darkness compared to the thickness condition, $t = 1.85$, $df = 121$, $p = .066$, estimate = 0.05, $SE = 0.03$, $d = .26$.
77. For each condition the slope was significantly less than 1 i.e., word condition: $t = -7.60$, $p < .001$, $d = 3.97$; darkness condition: $t = -8.52$, $p < .001$, $d = 3.62$; and thickness condition: $t = -13.01$, $p < .001$, $d = 4.24$.
78. For example, if a participant estimated the thickness of lines as .1, .2, .3, .4, .5, .6, .7, .8, .9, then their differences would all be .1. The *SD* of these differences would be approximately 0. In contrast, imagine another participant who estimated the lines as .1, .2, .2, .3, .4, .6, .6, .9, .9. The *SD* of these differences is .11.
79. The *SD*s were analyzed with a linear mixed model for which the fixed effect was visual condition and participant was the random effect. A comparison of means between the two visual conditions revealed a significant difference, $t = 2.04$, $p = .046$, $d = .31$.
80. Calculated as the *SD* for the thickness condition (0.121) divided by the sum for both conditions (0.104 + 0.121) and multiplied by 100, equals 54 per cent. See Figure 5 for these *SD*s by condition.
81. This result was obtained using a linear regression model with the range at each level of probability being communicated as the dependent measure and condition as the independent variable, $F(2, 21) = 1.38$, $p = .27$, $\varepsilon^2 = .03$. Pairwise *t*-tests were used to compare means i.e., for word versus darkness: $t = 1.30$, $p = .22$, $d = .65$, word versus thickness: $t = 1.62$, $p = .13$, $d = .81$, and darkness versus thickness: $t = -0.24$, $p = .81$, $d = .12$. These effect sizes seem large but the confidence interval around the effect size is very large. For word versus darkness, the 95 per cent CI is [−.21, 1.48]. For word versus thickness, the 95 per cent CI around the effect size d is [−.07, 1.66]. For thickness versus darkness, it is [−.68, .92].
82. A logistic regression model was used to assess the significance of these differences. The dependent measure was whether the two estimates were the same, the fixed effect was condition, and participant was the random effect. Post-hoc comparisons showed that the difference between the word and thickness conditions was significant, $t = -2.58$, $p = .01$. The difference between the darkness and word conditions was not significant, $t = -1.66$, $p = .10$. The difference between the darkness and thickness conditions was also not significant, $t = 0.91$, $p = .36$.
83. A linear mixed model with condition as the fixed effect and participant as the random effect was used to analyze the absolute difference scores. In addition, for comparison of means between the word and thickness conditions ($t = 2.58$, $df = 666$, $p = .01$, $d = .23$), for comparison between the word and darkness conditions ($t = 1.66$, $df = 665$, $p = .10$, $d = .15$), and for comparison between the darkness and thickness conditions, $t = -0.91$, $df = 668$, $p = .36$, $d = .08$.
84. To test whether these improvements were statistically significant, we ran *t*-tests between slopes for first and second evaluations. For both visual conditions, *p*s > .57, *d*s < .04.
85. For the word condition, $p = .081$, $d = .04$.
86. Dhami & Mandel, "Communicating Uncertainty."
87. see Dhami & Mandel, "Words or Numbers?"
88. Dhami, "Towards an Evidence-Based Approach"; Ho et al., "Improving the Communication"; Mandel, "Accuracy of Intelligence Forecasts"; Mandel & Irwin, "On Measuring Agreement"; Mandel & Irwin, "Tracking Accuracy"; and Wintle et al., "Verbal Probabilities."
89. see Dhami & Mandel, "Words or Numbers?"
90. Lavine & Gouin, *Applicability*; McCue, "Security Threats"; and Phillips et al., "Applying Social Network Analysis."
91. e.g. Weiss, "Communicating Uncertainty."

92. Hsee, “The Evaluability Hypothesis”; Hsee et al., “Preference Reversals”; and Hsee & Zhang, “General Evaluability Theory.”
93. e.g. Holliman et al., *Visual Entropy;* Kunze et al., “Augmented Reality Displays.”
94. e.g. Bisantz et al.,“Visual Representations”; Kubíček & Šašinka, “Thematic Uncertainty Visualization Usability”; and MacEachren et al., “Visual Semiotics.”
95. e.g. Beyth-Marom, “How Probable is Probable?”; Budescu et al., “The Interpretation”; Clarke et al., “Ratings of Verbal Expressions”; Dhami, & Wallsten, “Interpersonal Comparison”; Lichtenstein & Newman, “Empirical Scaling”; and Wiles, Duffy & Neill, “The Numerical Translation”.
96. Dhami, “Towards an Evidence-Based Approach”; Ho et al., “Improving the Communication”; Mandel & Irwin, “On Measuring Agreement”; Mandel & Irwin, “Tracking Accuracy”; and Wintle et al., “Verbal Probabilities.”
97. Hullman et al., “In Pursuit of Error.”
98. Ibid.
99. e.g. Budescu, Weinberg & Wallsten, “Decisions”; Dhami, & Wallsten, “Interpersonal Comparison”; and Wallsten, Fillenbaum & Cox, “Base-Rate Effects.”
100. e.g. Leitner & Buttenfield, “Guidelines.”
101. e.g. Kunze et al., “Augmented Reality Displays.”
102. Drecki, “Visualization of Uncertainty.”
103. Szafir, “Modeling Color Difference.”
104. Dhami, “Towards an Evidence-Based Approach”; Ho et al., “Improving the Communication”; and Mandel, “Accuracy of Intelligence Forecasts.”
105. Mandel & Irwin, “On Measuring Agreement”; Mandel & Irwin, “Tracking Accuracy”; and Wintle et al., “Verbal Probabilities.”
106. Budescu et al., “The Interpretation”; and Harris et al., “Lost in Translation?”
107. Ho et al., “Improving the Communication”; and See also Dhami, “Towards an Evidence-Based Approach.”
108. Dhami, “Towards an Evidence-Based Approach”; and Barnes, “Making Intelligence Analysis,” 11.
109. Dhami, “Towards an Evidence-Based Approach.”
110. Mosteller & Youtz, “Quantifying Probabilistic Expressions.”
111. Dhami, “Towards an Evidence-Based Approach”; and Willems, Albers & Smeets, “Variability.”
112. Renooij & Witteman, “Talking Probabilities.”
113. Budescu et al., “The Interpretation”; and Wintle et al., “Verbal Probabilities.”
114. Dhami, “Towards an Evidence-Based Approach”; Ho et al., “Improving the Communication”; Mandel, “Accuracy of Intelligence Forecasts”; Mandel & Irwin, “On Measuring Agreement”; Mandel & Irwin, “Tracking Accuracy”; and Wintle et al., “Verbal Probabilities.”
115. Willems, Albers & Smeets, “Variability.”
116. Ho et al., “Improving the Communication.”
117. Dhami et al. “Improving Intelligence Analysis.”
118. See Hullman et al., “In Pursuit of Error”; Jena et al.,“Uncertainty Visualisation”; and Kinkeldey, MacEachren & Schiewe “How to Assess.”
119. e.g. Dhami, & Wallsten, “Interpersonal Comparison.”
120. e.g. see Mandel & Irwin, “On Measuring Agreement”; and Wintle et al., “Verbal Probabilities.”
121. See Dhami & Mandel, “Communicating Uncertainty.”
122. See Padilla, Kay & Hullman, *Uncertainty Visualization*.
123. Edwards & Nelson, “Visualizing Data Certainty.”
124. e.g. Brun & Teigen, “Verbal Probabilities”; Cohen & Wallsten, “The Effect of Constant Outcome Value”; Harris & Corner, “Communicating Environmental Risks”; and Weber & Hilton, “Contextual Effects.”
125. Mandel, “Accuracy of Intelligence Forecasts”; and Mellers et al., “How Generalizable is Good Judgment?”
126. See Dhami & Mandel, “Words or Numbers?”
127. Knapp, Gardner & Woolf, “Combined Verbal and Numerical Expressions”; and Thompson et al., “Perceived Strength.”
128. e.g. Cheong et al., “Evaluating the Impact”; and McDowell & Kause, “Communicating Uncertainties.”

Acknowledgements

We would like to thank Velichka Hadzhieva for her research assistance.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Science Foundation grant numbers [BCS-1632222 and SES-2030059 to JKW].

Notes on contributors

Mandeep K. Dhami, PhD, is Professor in Psychology at Middlesex University, UK. Her research focuses on human decision-making; primarily in the law enforcement, justice, and security sectors. She has authored over 140 scholarly publications and is lead editor of 'Judgment and decision-making as a skill: Learning, development, and evolution' (Cambridge University Press, 2011). Mandeep has worked as a Defence Scientist in the UK, and regularly advises on matters related to intelligence analysis.

Jessica K. Witt, PhD, is Professor in Psychology at Colorado State University. She studies how to help people make decisions (especially using visualizations) when the outcomes are unknown or the information is uncertain. She has published extensively on this topic and her work is regularly covered by media outlets.

Peter De Werd is Associate Professor in intelligence and security at the Netherlands Defence Academy. Current research projects focus on publicly available information and intelligence analysis, intelligence oversight practices, and critical intelligence studies. He is author of US Intelligence and Al Qaeda (Edinburgh University Press, 2020).

Bibliography

Barnes, A. "Making Intelligence Analysis More Intelligent: Using Numeric Probabilities." *Intelligence and National Security* 31, no. 3 (2016): 327–344. doi:10.1080/02684527.2014.994955.

Belton, K., and M. K. Dhami. "Cognitive Biases and Debiasing Relevant to Intelligence Analysis." In *Handbook on Bounded Rationality*, edited by R. Viale, 548–569. London: Routledge, 2021.

Beyth-Marom, R. "How Probable is Probable? A Numerical Translation of Verbal Probability Expressions." *Journal of Forecasting* 1, no. 3 (1982): 257–269. doi:10.1002/for.3980010305.

Bisantz, A. M., R. T. Stone, J. Pfautz, A. Fouse, M. Farry, E. Roth, A. L. Nagy, and G. Thomas. "Visual Representations of Meta-Information." *Journal of Cognitive Engineering and Decision Making* 3 (2009): 67–91. doi:10.1518/155534309X433726.

Boukhelifa, N., A. Bezerianos, T. Isenberg, and J. D. Fekete. "Evaluating Sketchiness as a Visual Variable for the Depiction of Qualitative Uncertainty." *IEEE Transactions on Visualization and Computer Graphics* 18, no. 12 (2012): 2769–2778. doi:10.1109/TVCG.2012.220.

Brun, W., and K. H. Teigen. "Verbal Probabilities: Ambiguous, Context Dependent, or Both?" *Organizational Behavior and Human Decision Processes* 41, no. 3 (1988): 390–404. doi:10.1016/0749-5978(88)90036-2.

Budescu, D. V., H. H. Por, and S. B. Broomell. "Effective Communication of Uncertainty in the IPCC Reports." *Climatic Change* 113, no. 2 (2012): 181–200. doi:10.1007/s10584-011-0330-3.

Budescu, D. V., H. H. Por, S. B. Broomell, and M. Smithson. "The Interpretation of IPCC Probabilistic Statements Around the World." *Nature Climate Change* 4, no. 6 (2014): 508–512. doi:10.1038/nclimate2194.

Budescu, D. V., S. Weinberg, and T. S. Wallsten "Decisions Based on Numerically and Verbally Expressed Uncertainties." *Journal of Experimental Psychology Human Perception and Performance* 14, no. 2 (1988): 281–294. doi:10.1037/0096-1523.14.2.281.

Butler, C., J. Marshall, M. Mates, and A. Taylor. *Review of Intelligence on Weapons of Mass Destruction: Implementation of Its Conclusions*. 2004. https://fas.org/irp/world/uk/butler071404.pdf.

Cheong, L., S. Bleisch, A. Kealy, K. Tolhurst, T. Wilkening, and M. Duckham. "Evaluating the Impact of Visualization of Wildfire Hazard Upon Decision-Making Under Uncertainty." *International Journal of Geographical Information Science* 30, no. 7 (2016): 1377–1404. doi:10.1080/13658816.2015.1131829.

Chilcot, J. *The Report of the Iraq Inquiry. Executive Summary*. 2016. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/535407/The_Report_of_the_Iraq_Inquiry_-_Executive_Summary.pdf.

Chung, J., and S. Wark. "Visualising Uncertainty for Decision Support." DST-Group-TR-3325, 2016. https://www.dst.defence.gov.au/sites/default/files/publications/documents/DST-Group-TR-3325_0.pdf.

Clarke, V. A., C. L. Ruffin, D. J. Hill, and A. L. Beaman "Ratings of Orally Presented Verbal Expressions of Probability by a Heterogeneous Sample." *Journal of Applied Social Psychology* 22, no. 8 (1992): 638–656. doi:10.1111/j.1559-1816.1992.tb00995.x.

Cohen, B. L., and T. S. Wallsten. "The Effect of Constant Outcome Value on Judgments and Decision Making Given Linguistic Probabilities." *Journal of Behavioral Decision Making* 5 (1992): 53–72. doi:10.1002/bdm.3960050107.

College of Policing. *Delivering Effective Analysis*. n.d. https://www.app.college.police.uk/app-content/intelligence-management/analysis/delivering-effective-analysis/.

Collins, R. N., and D. R. Mandel. “Cultivating Credibility with Probability Words and Numbers.” *Judgment & Decision Making* 14, no. 6 (2019): 683–695. http://journal.sjdm.org/19/190912/jdm190912.html.

Correll, M., M. Moritz, and J. Heer. “Value-suppressing uncertainty palettes.” *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Paper 642. 1–11. New York, NY, Association for Computing Machinery, 2018. doi:10.1145/3173574.3174216.

DeCarlo, L. T. “Signal Detection Theory and Generalized Linear Models.” *Psychological Methods* 3, no. 2 (1998): 186–205. doi:10.1037/1082-989X.3.2.186.

Deitrick, S., and E. A. Wentz. “Developing Implicit Uncertainty Visualization Methods Motivated by Theories in Decision Science.” *Annals of the Association of American Geographers* 105, no. 3 (2015): 531–551. doi:10.1080/00045608.2015.1012635.

Dhami, M. K. “Towards an Evidence-Based Approach to Communicating Uncertainty in Intelligence Analysis.” *Intelligence and National Security* 33, no. 2 (2018): 257–272. doi:10.1080/02684527.2017.1394252.

Dhami, M. K., I. Belton, and K. Careless. “Critical Review of Analytic Techniques.” *Proceedings of the 2016 European Intelligence and Security Informatics Conference*, 152–155. Uppsala, Sweden, August 17–19, 2016. doi:10.1109/EISIC.2016.33.

Dhami, M. K., and D. R. Mandel. “Words or Numbers? Communicating Probability in Intelligence Analysis.” *The American Psychologist* 76, no. 3 (2021): 549–560. doi:10.1037/amp0000637.

Dhami, M. K., and D. R. Mandel “Communicating Uncertainty Using Words and Numbers.” *Trends in Cognitive Sciences* 26 (2022): 514–526. doi:10.1016/j.tics.2022.03.002.

Dhami, M. K., D. R. Mandel, B. A. Mellers, and P. E. Tetlock. “Improving Intelligence Analysis with Decision Science.” *Perspectives on Psychological Science* 10 (2015): 753–757. doi:10.1177/1745691615598511.

Dhami, M. K., and T. S. Wallsten “Interpersonal Comparison of Subjective Probabilities: Toward Translating Linguistic Probabilities.” *Memory & Cognition* 33, no. 6 (2005): 1057–1068. doi:10.3758/BF03193213.

Dong, M., L. Chen, L. Wang, X. Jiang, and G. Chen. “Uncertainty Visualization for Mobile and Wearable Devices Based Activity Recognition Systems.” *International Journal of Human–Computer Interaction* 33, no. 2 (2017): 151–163. doi:10.1080/10447318.2016.1224527.

Drecki, I. “Visualization of Uncertainty in Geographical Data.” In *Spatial Data Quality*, edited by W. Shi, P. F. Fisher, and M. F. Goodchild, 140–159. London: Taylor & Francis, 2002.

Edwards, L. D., and E. S. Nelson. “Visualizing Data Certainty: A Case Study Using Graduated Circle Maps.” *Cartographic Perspectives* 38 (2001): 19–36. doi:10.14714/CP38.793.

Fernandes, M., L. Walls, S. Munson, J. Hullman, and M. Kay. “Uncertainty Displays Using Quantile Dotplots or Cdfs Improve Transit Decision-Making.” *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–12. 2018. doi:10.1145/3173574.3173718.

Friedman, J. A., J. D. Baker, B. A. Mellers, P. E. Tetlock, and R. Zeckhauser. “The Value of Precision in Probability Assessment: Evidence from a Large-Scale Geopolitical Forecasting Tournament.” *International Studies Quarterly* 62, no. 2 (2018): 410–422. doi:10.1093/isq/sqx078.

Friedman, J. A., and R. Zeckhauser. “Assessing Uncertainty in Intelligence.” *Intelligence and National Security* 27, no. 6 (2012): 824–847. doi:10.1080/02684527.2012.708275.

Garcia-Retamero, R., Y. Okan, and E. T. Cokely. “Using Visual Aids to Improve Communication of Risks About Health: A Review.” *Scientific World Journal* 2012 (2012): 1–10. doi:10.1100/2012/562637.

Gentry, J. A., and J. S. Gordon. *Strategic Warning Intelligence: History, Challenges, and Prospects*. Georgetown: Georgetown University Press, 2019.

Harris, A. J. L., and A. Corner. “Communicating Environmental Risks: Clarifying the Severity Effect in Interpretations of Verbal Probability Expressions.” *Journal of Experimental Psychology: Learning, Memory & Cognition* 37, no. 6 (2011): 1571–1578. doi:10.1037/a0024195.

Harris, A. J. L., A. Corner, J. Xu, and X. Du. “Lost in Translation? Interpretations of the Probability Phrases Used by the Intergovernmental Panel on Climate Change in China and the U.K.” *Climatic Change* 121, no. 2 (2013): 415–425. doi:10.1007/s10584-013-0975-1.

Hart, A., L. Maxim, M. Siegrist, N. V. Goetz, C. da Cruz, C. Merten, O. Mosbach-Schulz, M. Lahaniatis, A. Smith, and A. Hardy. “Guidance on Communication of Uncertainty in Scientific Assessments. European Food Safety Authority.” *EFSA Journal* 17, no. 1 (2019): 1–73. doi:10.2903/j.efsa.2019.5520.

Ho, E. H., D. V. Budescu, M. K. Dhami, and D. R. Mandel. “Improving the Communication of Uncertainty in Climate Science and Intelligence Analysis.” *Behavioral Science and Policy* 1, no. 2 (2015): 43–55. doi:10.1177/237946151500100206.

Hogan Carr, R., B. Montz, K. Maxfield, S. Hoekstra, K. Semmens, and E. Goldman. “Effectively Communicating Risk and Uncertainty to the Public: Assessing the National Weather service’s Flood Forecast and Warning Tools.” *Bulletin of the American Meteorological Society* 97, no. 9 (2016): 1649–1665. https://journals.ametsoc.org/view/journals/bams/97/9/bams-d-14-00248.1.xml.

Holliman, N., S. Fernstad, M. Simpson, and K. Wilson. *Visual Entropy and the Visualization of Uncertainty*. 2019. 10.48550/arXiv.1907.12879.

Hsee, C. K. “The Evaluability Hypothesis: An Explanation for Preference Reversals Between Joint and Separate Evaluations of Alternatives.” *Organizational Behavior and Human Decision Processes* 67, no. 3 (1996): 247–257. doi:10.1006/obhd.1996.0077.

Hsee, C. K., G. F. Loewenstein, S. Blount, and M. H. Bazerman. “Preference Reversals Between Joint and Separate Evaluations of Options: A Review and Theoretical Analysis.” *Psychological Bulletin* 125, no. 5 (1999): 576–590. doi:10.1037/0033-2909.125.5.576.

Hsee, C. K., and J. Zhang. “General Evaluability Theory.” *Perspectives on Psychological Science* 5, no. 4 (2010): 343–355. doi:10.1177/1745691610374586.

Hullman, J., X. Qiao, M. Correll, A. Kale, and M. Kay. “In Pursuit of Error: A Survey of Uncertainty Visualization Evaluation.” *IEEE Transactions on Visualization and Computer Graphics* 25, no. 1 (2019): 903–913. doi:10.1109/TVCG.2018.2864889.

Janzwood, S. “Confident, Likely, or Both? The Implementation of the Uncertainty Language Framework in IPCC Special Reports.” *Climatic Change* 162, no. 3 (2020): 1655–1675. doi:10.1007/s10584-020-02746-x.

Jena, A., U. Engelke, T. Dwyer, V. Rajamanickam, and C. Paris. “Uncertainty Visualisation: An Interactive Visual Survey.” Paper Presented at the IEEE Pacific Visualization SymposiumTianjin, June 3–5, 2020. doi:10.1109/PacificVis48177.2020.1014.

Johnson, E. M. *Numerical Encoding of Qualitative Expressions of Uncertainty*. Arlington, VA: US Army Research Institute for the Behavioral and Social Sciences, 1973.

Juanchich, M., and M. Sirota. “Do People Really Prefer Verbal Probabilities?” *Psychological Research* 84, no. 8 (2020): 2325–2338. doi:10.1007/s00426-019-01207-0.

Kent, S. *Words of Estimative Probability*. 1964. https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/books-and-monographs/sherman-kent-and-the-board-of-national-estimates-collected-essays/6words.html.

Kesselman, R. F. *Verbal Probability Expressions in National Intelligence Estimates: A Comprehensive Analysis of Trends from the Fifties Through Post 9/11*, Unpublished Masters diss. Erie, PA: Mercyhurst College, 2008. https://gwern.net/doc/statistics/bayes/2008-kesselman.pdf.

Kinkeldey, C., A. M. MacEachren, M. Riveiro, and J. Schiewe. “Evaluating the Effect of Visually Represented Geodata Uncertainty on Decision-Making: Systematic Review, Lessons Learned, and Recommendations.” *Cartography and Geographic Information Science* 44, no. 1 (2017): 1–21. doi:10.1080/15230406.2015.1089792.

Kinkeldey, C., A. M. MacEachren, and J. Schiewe. “How to Assess Visual Communication of Uncertainty? A Systematic Review of Geospatial Uncertainty Visualisation User Studies.” *The Cartographic Journal* 51, no. 4 (2014): 372–386. doi:10.1179/1743277414Y.0000000099.

Knapp, P., P. H. Gardner, and E. Woolf. “Combined Verbal and Numerical Expressions Increase Perceived Risk of Medicine Side-Effects: A Randomized Controlled Trial of EMA Recommendations.” *Health Expectations* 19 (2016): 264–274. doi:10.1111/hex.12344.

Kolesnik, K., S. Silska-Gembkaa, and J. Gierusz. “The Interpretation of the Verbal Probability Expressions Used in the IFRS – the Differences Observed Between Polish and British Accounting Professionals.” *Journal of Accounting and Management Information Systems* 18, no. 1 (2019): 25–49. doi:10.24818/jamis.2019.01002.

Korporaal, M., I. T. Ruginski, and S. I. Fabrikant. “Effects of Uncertainty Visualization on Decision Making with Map-Based Geographic Data Under Time Pressure.” *Frontiers in Computer Science* 2 (2020): 32. doi:10.3389/fcomp.2020.00032.

Kubíček, P., and C. Šašinka. “Thematic Uncertainty Visualization Usability – Comparison of Basic Methods.” *Annals of GIS* 17, no. 4 (2011): 253–263. doi:10.1080/19475683.2011.625978.

Kunze, A., S. J. Summerskill, R. Marshall, and A. J. Filtness. “Augmented Reality Displays for Communicating Uncertainty Information in Automated Driving.” *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomativeUI’18)*., 164–175 New York, NY, Association for Computing Machinery, 2018. doi:10.1145/3239060.3239074.

Kuznetsova, A., P. B. Brockhoff, and R. H. B. Christensen. “lmerTest Package: Tests in Linear Mixed Effects Models.” *Journal of Statistical Software* 82, no. 13 (2017): 1–26. doi:10.18637/jss.v082.i13.

Larkin, J. H., and H. A. Simon. “Why a Diagram is (Sometimes) Worth ten Thousand Words.” *Cognitive Science* 11 (1987): 65–100. doi:10.1111/j.1551-6708.1987.tb00863.x.

Lavigne, V., and D. Gouin. *Applicability of Visual Analytics to Defence and Security Operations*. 2011. https://apps.dtic.mil/sti/pdfs/ADA546992.pdf.

Leitner, M., and B. P. Buttenfield. “Guidelines for the Display of Attribute Certainty.” *Cartography and Geographic Information Science* 27 (2000): 3–14. doi:10.1559/152304000783548037.

Lenth, R. “Emmeans: Estimated Marginal Means, Aka Least-Squares Means. R Package Version 1.10.3.” Accessed February 8, 2025. https://CRAN.R-project.org/package=emmeans.

Lichtenstein, S., and J. R. Newman. “Empirical Scaling of Common Verbal Phrases Associated with Numerical Probabilities.” *Psychonomic science* 9, no. 10 (1967): 563–564. doi:10.3758/BF03327890.

Lonsdale, D., and M. dos Santos Lonsdale. “Handling and Communicating Intelligence Information: A Conceptual, Historical and Information Design Analysis.” *Intelligence and National Security* 34, no. 5 (2019): 703–726. doi:10.1080/02684527.2019.1592841.

MacEachren, A. M., R. E. Roth, J. O'Brien, B. Li, D. Swingley, and M. Gahegan. "Visual Semiotics & Uncertainty Visualization: An Empirical Study." *IEEE Transactions on Visualization and Computer Graphics* 18, no. 12 (2012): 2496–2505. doi:10.1109/TVCG.2012.279.

Mandel, D. R. "Accuracy of Intelligence Forecasts from the Intelligence consumer's Perspective." *Policy Insights from the Behavioral and Brain Sciences* 2, no. 1 (2015): 111–120. doi:10.1177/2372732215602907.

Mandel, D. R., M. K. Dhami, S. Tran, and D. Irwin. "Arithmetic Computation with Probability Words and Numbers." *Journal of Behavioral Decision Making* 34, no. 4 (2021): 593–608. doi:10.1002/bdm.2232.

Mandel, D. R., and D. Irwin. "Facilitating Sender-Receiver Agreement in Communicated Probabilities: Is it Best to Use Words, Numbers or Both?" *Judgment & Decision Making* 16, no. 2 (2021a): 363–393. doi:10.1017/S1930297500008603.

Mandel, D. R., and D. Irwin. "Tracking Accuracy of Strategic Intelligence Forecasts: Findings from a Long-Term Canadian Study." *Futures and Foresight Science* 3, no. 3–4 (2021b): 1–13. doi:10.1002/ffo2.98.

Mandel, D. R., D. Irwin, and D. Pamucar. "On Measuring Agreement with Numerically Bounded Linguistic Probability Schemes: A Re-Analysis of Data from Wintle, Fraser, Wills, Nicholson, and Fidler (2019)." *PLoS One* 16, no. 3 (2021): e0248424. doi:10.1371/journal.pone.0248424.

Marchio, J. ""If the Weatherman can. . .": The Intelligence community's Struggle to Express Analytic Uncertainty in the 1970s." *Studies in Intelligence* 58, no. 4 (2014): 31–42.

McCue, C. "Security Threats, Humanitarian Needs: Operations Research, Geospatial Analysis Boost Public Safety and Humanitarian Relief Efforts." *OR/MS Today* 41, no. 6 (2014). https://link.gale.com/apps/doc/A419534217/AONE?u=anon~808c7f20&sid=googleScholar&xid=d678caa9.

McDowell, M., and A. Kause. "Communicating Uncertainties About the Effects of Medical Interventions Using Different Display Formats." *Risk Analysis* 41 (2021): 2220–2239. doi:10.1111/risa.13739.

McKenzie, G., M. Hegarty, T. Barrett, and M. Goodchild. "Assessing the Effectiveness of Different Visualizations for Judgments of Positional Uncertainty." *International Journal of Geographical Information Science* 30, no. 2 (2016): 221–239. doi:10.1080/13658816.2015.1082566.

Mellers, B. A., J. D. Baker, E. Chen, D. R. Mandel, and P. E. Tetlock. "How Generalizable is Good Judgment? A Multi-Task, Multibenchmark Study." *Judgment & Decision Making* 12, no. 4 (2017): 369–381. http://journal.sjdm.org/17/17408/jdm17408.pdf.

Milne, A. E., M. J. Glendining, R. M. Lark, S. A. M. Perryman, T. Gordon, and A. P. Whitmore. "Communicating the Uncertainty in Estimated Greenhouse Gas Emissions from Agriculture." *Journal of Environmental Management* 160 (2015): 139–153. doi:10.1016/j.jenvman.2015.05.034.

Mosteller, F., and C. Youtz. "Quantifying Probabilistic Expressions." *Statistical Science* 5, no. 1 (1990): 2–34. doi:10.1214/ss/1177012242.

NATO Standardization Office. *AJP-2.1, Edition B, Version 1: Allied Joint Doctrine for Intelligence Procedures*. Brussels: NATO, 2016.

Padilla, L., M. Kay, and J. Hullman. "Uncertainty Visualization." In Computational Statistics in Data Science, edited by W. Piegorsch, R. Levine, H. Zhang, and T. Lee, 405–421. Hoboken: Wiley, 2020. 10.31234/osf.io/ebd6r.

Phillips, E., J. R. C. Nurse, M. Goldsmith, and S. Creese. "Applying Social Network Analysis to Security." Paper Presented at the International Conference on Cyber Security for Sustainable Society, 11–27, Coventry, February 26, 2015. http://www.cs.ox.ac.uk/files/7193/csss2015_phillips_et_al.pdf.

Phillips, E., J. R. C. Nurse, M. Goldsmith, and S. Creese. "Applying social network analysis to security." *International Conference on Cyber Security for Sustainable Society*, 1127. http://www.cs.ox.ac.uk/files/7193/csss2015_phillips_et_al.pdf.

R Core Team. "R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing." Accessed February 8, 2025. https://www.R-project.org/.

Renooij, S., and C. Witteman. "Talking Probabilities: Communicating Probabilistic Information with Words and Numbers." *International Journal of Approximate Reasoning* 22 (1999): 169–194. doi:10.1016/S0888-613X(99)00027-4 .

Reyna, V. F., and C. J. Brainerd. "Numeracy, Ratio Bias, and Denominator Neglect in Judgments of Risk and Probability." *Learning & Individual Differences* 18 (2008): 89–107. doi:10.1016/j.lindif.2007.03.011.

Sanyal, J., S. Zhang, G. Bhattacharya, P. Amburn, and R. Moorhead. "A User Study to Compare Four Uncertainty Visualization Methods for 1D and 2D Datasets." *IEEE Transactions on Visualization and Computer Graphics* 15 (2009): 1209–1218. doi:10.1109/TVCG.2009.114.

Singh, P. P., and D. Philip. "An Innovative Color-Coding Scheme for Terrorism Threat Advisory System." *Methodological Innovations* 16, no. 1 (2023): 38–56. doi:10.1177/20597991221144577.

Spiegelhalter, D., M. Pearson, and I. Short. "Visualizing Uncertainty About the Future." *Science* 333, no. 6048 (2011): 1393–1400. doi:10.1126/science.1191181.

Szafir, D. A. "Modeling Color Difference for Visualization Design." *IEEE Transactions on Visualization and Computer Graphics* 24, no. 1 (2018): 392–401. doi:10.1109/TVCG.2017.2744359.

Taylor, A. L., S. Dessai, and W. Bruine de Bruin. "Communicating Uncertainty in Seasonal and Interannual Climate Forecasts in Europe." *Philosophical Transactions* 373, no. 2055 (2015): 20140454. doi:10.1098/rsta.2014.0454(2015).

Teigen, K. H., and W. Brun. "Yes, but it is Uncertain: Direction and Communicative Intention of Verbal Probability Terms." *Acta Psychologica* 88, no. 3 (1995): 233–258. doi:10.1016/0001-6918(93)E0071-9.

Teigen, K. H., and W. Brun. “The Directionality of Verbal Probability Expressions: Effects on Decisions, Predictions, and Probabilistic Reasoning.” *Organizational Behavior and Human Decision Processes* 80, no. 2 (1999): 155–190. doi:10.1006/obhd.1999.2857.

Teigen, K. H., M. Juanchich, and E. Løhre. “Combining Verbal Forecasts: The Role of Directionality and the Reinforcement Effect.” *Journal of Behavioral Decision Making* 36, no. 2 (2022). doi:10.1002/bdm.2298.

Thompson, W. C., R. Hofstein Grady, E. Lai, and H. S. Stern. “Perceived Strength of Forensic scientists’ Reporting Statements About Source Conclusions.” *Law, Probability and Risk* 17, no. 2 (2018): 133–155. doi:10.1093/lpr/mgy012.

Turkay, C., T. von Landesberger, D. Archambault, S. Liu, and R. Chang. “Special Issue on Interactive Visual Analytics for Making Explainable and Accountable Decisions.” *ACM Transactions on Interactive Intelligent Systems* 11, no. 3–4 (2021): 1–4. doi:10.1145/3471903.

U.S. Central Intelligence Agency. *Report on a Study of Intelligence Judgments Preceding Significant Historical Failures: The Hazards of Single-Outcome Forecasting*. Washington, DC: CIA, 1983. https://www.cia.gov/library/readingroom/docs/CIA-RDP86B00269R001100100010-7.pdf.

U.S. Congressional Select Committee on Intelligence. “Report on the U.S.” In Intelligence Community’s Prewar Intelligence Assessments on Iraq, Washington, DC: US Congress, 2005. https://fas.org/irp/congress/2004_rpt/ssci_iraq.pdf.

U.S. Office of the Director of National Intelligence. *Intelligence Community Directive 203, Analytic Standards*. Washington, DC: US ODNI, 2015. https://fas.org/irp/dni/icd/icd-203.pdf .

U.S. Office of the Director of National Intelligence (US ODNI). *Prospects for Iraq’s Stability: A Challenging Road Ahead. National Intelligence Estimate*. Washington, DC: US ODNI, 2007. https://irp.fas.org/dni/iraq020207.pdf.

Wallsten, T. S., D. V. Budescu, R. Zwick, and S. M. Kemp. “Preferences and Reasons for Communicating Probabilistic Information in Verbal or Numerical Terms.” *Bulletin of the Psychonomic Society* 31, no. 2 (1993): 135–138. doi:10.3758/BF03334162.

Wallsten, T. S., S. Fillenbaum, and A. Cox. “Base-Rate Effects on the Interpretations of Probability and Frequency Expressions.” *Journal of Memory and Language* 25, no. 5 (1986): 571–587. doi:10.1016/0749-596X(86)90012-4.

Wallsten, T. S., Y. Shlomi, and H. Ting. “Exploring Intelligence analysts’ Selection and Interpretation of Probability Terms. Final Report for Research Contact ‘Expressing Probability in Intelligence analysis’.” *Sponsored by the CIA* (2008).

Wark, D. L. “The Definition of Some Estimative Expressions.” *Studies in Intelligence* 8, no. 4 (1964): 67–80.

Weber, E. U., and D. J. Hilton. “Contextual Effects in the Interpretations of Probability Words: Perceived Base Rate and Severity of Events.” *Journal of Experimental Psychology Human Perception and Performance* 16, no. 4 (1990): 781–789. doi:10.1037//0096-1523.16.4.781.

Weiss, C. “Communicating Uncertainty in Intelligence and Other Professions.” *International Journal of Intelligence & CounterIntelligence* 21, no. 1 (2008): 57–85. doi:10.1080/08850600701649312.

Westfall, J., D. A. Kenny, and C. M. Judd. “Statistical Power and Optimal Design in Experiments in which Samples of Participants Respond to Samples of Stimuli.” *Journal of Experimental Psychology General* 143, no. 5 (2014): 2020–2045. doi:10.1037/xge0000014.

Wiles, M. D., A. Duffy, and K. Neill. “The Numerical Translation of Verbal Probability Expressions by Patients and Clinicians in the Context of Peri-Operative Risk Communication.” *Anaesthesia* 75, no. 1 (2020): e39–e45. doi:10.1111/anae.14871.

Willems, S. J. W., C. J. Albers, and I. Smeets. “Variability in the Interpretation of Probability Phrases Used in Dutch News Articles – a Risk for Miscommunication.” *Journal of Science Communication* 19, no. 2 (2020): A03. doi:10.22323/2.19020203.

Wintle, B. C., H. Fraser, B. C. Wills, A. E. Nicholson, F. Fidler, and E. Yechiam. “Verbal Probabilities: Very Likely to Be Somewhat More Confusing Than Numbers.” *PLoS One* 14, no. 4 (2019): e0213522. doi:10.1371/journal.pone.0213522.

Zimmer, A. C. “Verbal Vs. Numerical Processing of Subjective Probabilities.” *Advances in Psychology* 16 (1983): 159–182. doi:10.1016/S0166-4115(08)62198-6.

Zimmer, A. C. “A Model for the Interpretation of Verbal Predictions.” *International Journal of Man-Machine Studies* 20, no. 1 (1984): 121–134. doi:10.1016/S0020-7373(84)80009-7.