Judgment and Decision Making, Vol. 16, No. 2, March 2021, pp. 363–393

Facilitating sender-receiver agreement in communicated probabilities: Is it best to use words, numbers or both?

David R. Mandel* Daniel Irwin†

Abstract

Organizations tasked with communicating expert judgments couched in uncertainty often use numerically bounded linguistic probability schemes to standardize the meaning of verbal probabilities. An experiment (N = 1,202) was conducted to ascertain whether agreement with such a scheme was better when probabilities were presented verbally, numerically or in a combined "verbal + numeric" format. Across three agreement measures, the numeric and combined formats outperformed the verbal format and also yielded better discrimination between low and high probabilities and were less susceptible to the fifty-fifty blip phenomenon. The combined format did not confer any advantage over the purely numeric format. The findings indicate that numerically bounded linguistic probability schemes are an ineffective means of communicating information about probabilities to others and they call into question recommendations for use of the combined format for delivering such schemes.

Keywords: verbal probability, numeric probability, uncertainty communication, agreement

1 Introduction

Expert judgments are frequently made under conditions of uncertainty. Consequently, those judgments are often conveyed as probability estimates to end-users whose decisions and outcomes, in turn, may be affected by such information. For instance, in medicine, the

*Defence Research and Development Canada, 1133 Sheppard Ave W., North York, ON M3K 2C9, Canada. Email: drmandel66@gmail.com. https://orcid.org/0000-0003-1036-2286.

†Department of National Defence

This research was supported by Canadian Safety and Security Program project CSSP-2018-TI-2394. We thank Robert Collins for assistance with plotting the figures. We thank Mandeep Dhami, Bonnie Wintle and an anonymous reviewer for comments on an earlier draft of this paper. Supplementary files are available from the Open Science Foundation at https://osf.io/eztf8/.

https://doi.org/10.1017/S1930297500008603 Published online by Cambridge University Press

estimated probability of adverse side effects may influence patients' willingness to undergo certain treatments (Berry et al., 1997; Ziegler et al., 2001). In criminal proceedings, expert communication of uncertainty underlying forensic evidence may shape the conclusions of judges or juries (Ligertwood & Edmond, 2012; McQuiston-Surret & Saks, 2008). In national security policymaking, intelligence assessments are usually qualified by probabilities that can shape consequential decisions, including whether to go to war (Kent, 1964; Marchio, 2014; Debs & Monteiro, 2014). Indeed, the communication of uncertainty is central to all domains of public policymaking (Funtowicz & Ravetz, 1990).

Many organizations and professional groupings that produce expert judgments prefer to express uncertainties with verbal probabilities such as "likely" or "unlikely" rather than with precise numeric probabilities such as "70% chance" or imprecise numeric ranges such as "60% to 80% chance" (e.g., Dhami & Mandel, 2020; Ho et al., 2015). For instance, in a recent study of National Weather Service tweets, 99.9% of probabilistic forecasts were made using verbal probability expressions (Lenhardt et al., 2020). Accountants also tend to prefer using verbal probabilities, despite the quantitative basis of the profession (Kolesnika et al., 2019). This tendency is partly attributable to the belief that end-users will not be able to effectively process numeric probabilities (e.g., Lewis et al., 2019) and it is partly attributable to the greater ease of producing assessments that are qualitatively rather than quantitatively qualified (Wallsten et al., 1993). As Beyth-Marom (1982) also suggested, the preference for using verbal probabilities may also be motivated by a desire to have one's probabilistic judgments remain less verifiable in terms of accuracy. Several studies have shown a "communication mode preference paradox" in which, on average, senders prefer verbal probabilities but receivers prefer numeric probabilities (Brun & Teigen, 1988; Erev & Cohen, 1990; Wallsten et al., 1993). In spite of senders' preferential tendency, extensive research has shown intrapersonal imprecision and interpersonal inconsistency in how people translate verbal probabilities into numeric equivalents (e.g., Beyth-Marom, 1982; Budescu & Wallsten, 1985; Dhami & Wallsten, 2005; Harris et al., 2013; Lichtenstein & Newman, 1967).

The detrimental consequences of using verbal probabilities to convey uncertainties have been noted in past literature (e.g., Dhami & Mandel, 2020; European Food Safety Authority et al., 2018; Friedman, 2019; Mandel & Irwin, 2020; Morgan, 1998). As noted already, verbal probabilities are fuzzy in their interpretation and can vary greatly in meaning across individuals. Compared to numeric probabilities, verbal probabilities are judged to be less clear in their communication of degrees of probability (Collins & Mandel, 2019), and verbal probabilities are prone to communicating implicit recommendations for action through their directionality (Teigen & Brun, 1995, 1999) — recommendations which may have policy-biasing effects in contexts such as national security intelligence, which has long focused on sustaining policy neutrality (Kent, 1951). However, instead of using numeric probabilities in their communications of risk and uncertainty, most organizations that disseminate probabilistic assessments have adopted numerically bounded linguistic probability (NBLP)

https://doi.org/10.1017/S1930297500008603 Published online by Cambridge University Press

schemes that prescribe an ordinal scale of verbal probabilities, each associated with numeric probability ranges (Ho et al., 2015; Mandel, Wallsten et al. 2021).[1] For example, Table 1 shows the five-point NBLP scheme currently used in NATO intelligence doctrine (2016; Dhami & Mandel, 2020, and Mandel & Irwin, 2020, discuss other NBLP schemes used in intelligence communities for communicating probabilities). According to this methodology, an analyst who judges an event to have a probability ≥ 60% and ≤ 90% should describe it as *likely*. Conversely, an analyst who describes an event as *likely* should agree that the probability falls within the associated range.

TABLE 1: NATO (2016) standard for communicating probability in intelligence.

Probability	
More than 90%	*Highly likely*
60% – 90%	*Likely*
40% – 60%	*Even chance*
10% – 40%	*Unlikely*
Less than 10%	*Highly unlikely*

However, studies show that even when participants are given the relevant NBLP scheme, they continue to show poor agreement with it (measured as the percentage of overlap between the numeric ranges in the standard and participants' ranges or by the proportion of participants whose best numeric equivalence estimates fall within the stipulated ranges). In a study on verbal probabilities used to communicate projections by the International Panel on Climate Change (IPCC), Budescu et al. (2009) asked participants to characterize the intended numeric meaning of each term (i.e., *very unlikely*, *unlikely*, *likely*, and *very likely*) by estimating its lower and upper bounds and a best estimate. The terms were embedded in sentences extracted from IPCC reports. Participants either received no guidance regarding the numeric equivalents of the verbal terms (control condition), unrestricted access to the IPCC translation table that contained numerical equivalents (translation condition), or numeric equivalents embedded in the sentences alongside each verbal probability (combined condition; e.g., "very likely [90% chance or greater]"). The combined format yielded better agreement than the translation or control formats. Median responses for the expressions were also less regressive and interpreted ranges were significantly narrower in the combined condition than in the other conditions. Subsequent replications including one with samples taken from 24 countries and in 17 languages also found better performance of the combined format (Budescu et al., 2012, 2014), and the better performance of the combined format was also generalized to a different standard used by the US intelligence community in Wintle et al. (2019) and in a re-analysis of the same dataset using a different agreement measure (Mandel & Irwin, 2021).

[1]We use the terms *verbal probability* and *linguistic probability* interchangeably.

https://doi.org/10.1017/S1930297500008603 Published online by Cambridge University Press

1.1 The present research

Our research expands on previous studies examining agreement between receivers' interpretations of verbal probability terms and the stipulated meaning of such terms in probability lexicons in multiple respects. First, a question yet to be investigated is whether agreement is affected by the presence or absence of numeric probability ranges in NBLP schemes — that is, does it help to numerically bound the verbal terms used in such schemes? Whereas previous studies (e.g., Budescu et al., 2009, 2014; Wintle et al., 2019) have examined the effect of introducing numeric ranges alongside verbal probability terms in specific assessments (i.e., the combined format), none of these studies examine schemes that themselves lack numeric bounds on the prescribed set of terms. This issue is important, however, because some probability schemes do not include probability ranges but merely consist of an ordered set of probability terms. Such approaches are common in risk assessment where an ordinal scale of probability terms is crossed with an ordinal scale of consequence severity to yield a risk matrix (Friedman, 2019; Mandel, 2007). To address this issue, we manipulated whether participants presented with the NATO lexicon shown earlier received the full version (as in Table 1) or a partial version that omitted the numeric ranges. Although agreement has been shown to be low even where numeric ranges are included in lexicons, we propose Hypothesis 1: agreement will be lower when numeric ranges are omitted from the probability scheme than when they are included.

A second aim of the present research was to build on studies by Budescu et al. (2009, 2012, 2014) and Wintle et al. (2019) by including a purely numeric probability format condition in which, following exposure to the full NATO lexicon, the intelligence assessment conveyed probabilities with numeric probability ranges only. We know of only one study that directly compared the effect of a combined probability format to a numeric format. In that study, Knapp et al. (2016) compared the effect of presenting information about the risk of a cancer medical treatment using either verbal expressions of relative frequency (e.g., "common") paired with upper-bounded numeric quantifiers (e.g., "up to 1 in 10") or using only the numeric quantifiers. Participants tended to overestimate risks in both conditions, but the degree of overestimation was far greater in the combined condition. Participants' judgments were also more variable in the combined condition than in the numeric condition. Knapp et al.'s (2016) findings call into question the benefit of pairing numeric expressions of probability with verbal probabilities. Unlike Knapp et al., we compared the agreement yielded by combined and numeric formats. Given that the combined format creates an opportunity for conflict between two sources of probability information, we propose Hypothesis 2: agreement using the numeric format will be as good as that observed using the combined format, and these formats (numeric and combined) will each show better agreement than the verbal format. If agreement levels were found to be as good or better when using the numeric range format, it would call into question why various organizations and professional groups remain committed to expressing probabilities primarily with verbal probabilities.

https://doi.org/10.1017/S1930297500008603 Published online by Cambridge University Press

A third aim of our research was to compare agreement using three distinct measures. As in past studies (Budescu et al., 2009; Wintle et al., 2019), we used the proportion of participants who provided "best-estimate" numeric equivalents that fell within the numeric ranges stipulated in the NATO standard. This measure captures an "all or none" interpretation of agreement. An alternative measure we tested that also uses best estimates measures agreement as the absolute distance between a participant's best estimate and the midpoint of the stipulated numeric range for a given verbal probability term. Under a variety of distributional assumptions (e.g., normal, rectangular or other symmetric distributions), the expected value of a numeric range is its midpoint. In interval analysis (Moore et al., 2009), for instance, a range is equivalent to its midpoint with a margin of error equal to half the range. We therefore included this measure, which had not been used in earlier agreement studies by Budescu et al. (2009, 2012, 2014) and Wintle et al. (2019). Finally, we computed an agreement measure that uses participants' upper and lower bounds in calculating the percentage overlap with a stipulated range, as shown in Equation 1:

$$PO = \frac{1 - \left[\max(U_e - U_s, 0) + \max(L_e - L_s, 0)\right]}{U_e - L_e} \times 100, \tag{1}$$

where L_e and U_e refer to the participant's lower-bound and upper-bound estimates, respectively, and U_s and L_s refer to the relevant lower and upper bounds stipulated in the NATO scheme. Wintle et al. (2019) also used a measure of percentage overlap, as shown in Equation 2:

$$PO_W = \frac{\min(U_e, U_s) - \max(L_e, L_s)}{\max(U_e, U_s) - \min(L_e, L_s)} \times 100 \tag{2}$$

However, unlike our measure, their measure penalizes in-range precision. For instance, if a participant provided lower and upper bounds of 45% and 55% for the term *even chance*, using Equation 2, the participant would be said to have 50% overlap of the stipulated 40%-60% range for this term (see Table 1). In contrast, our measure would score this participant as showing 100% overlap because 100% of their range was within the bounds of the stipulated range. In other words, the measure used in the present research does not punish within-range precision. Mandel and Irwin (2021) re-analyzed data from Wintle et al. (2019) using

the new percentage overlap measure and found that, as expected, agreement was higher across format conditions. However, the effect of format was not influenced by the choice of measure. In the present experiment, we hypothesized that the effect of probability format on agreement specified in Hypothesis 2 will be upheld across the three measures (Hypothesis 3).

A fourth aim of the present research was to examine whether interpretations of probability assessments are affected by the semantic context of the events they describe. Several context effects on the interpretation of verbal probabilities have been reported (Brun & Teigen, 1988; Mellers et al., 2017; Wallsten et al., 1986; Weber & Hilton, 1990), as well as on the selection of verbal probabilities from the sender's perspective (e.g., Patt & Schrag,

https://doi.org/10.1017/S1930297500008603 Published online by Cambridge University Press

2003). Some studies have found that interpretations of probability terms depend on the valence of the events qualified by such terms (e.g., Mullet & Rivet, 1991). In one study (Mandel, 2015a), participants discriminated better among the meanings of verbal probability terms ranging from extremes of *will not* (i.e., a very low probability) to *will* (a very high probability) when the terms referred to an event success rather than an event failure (and in both cases the desirability of the event was opaque, making it difficult to judge whether success or failure was "a good thing"). More recently, consistent with these findings, Dhami and Mandel (2021) found that participants considering a forensic assessment case discriminated better between the terms *probable* and *improbable* when the context was positive (i.e., the defendant was judged as being fit to plead) rather than negative (i.e., the defendant was judged as being *not* fit to plead).

In the present research, we attempted to generalize this valence-discrimination relation to a task in which valence was manipulated through gain/loss *framing* of outcomes rather than a valence-based *reflection* of outcomes (Fagley, 1993). Specifically, probabilistic assessments focused on the potential outcome of either saving half the lives of 1,000 threatened people (i.e., the positive frame) or losing half the lives of 1,000 threatened people (i.e., the negative frame). Consistent with the findings of Mandel (2015a) and Dhami and Mandel (2021), we tested Hypothesis 4: discrimination between the terms *unlikely* and *likely* (operationalized as the mean difference in numeric probability equivalents assigned to these terms) will be significantly better in the positive frame than in the negative frame. If so, this result should be expressed as a three-way interaction between probability format, probability level and frame in which differential discriminability (characterized by a probability level × frame interaction effect) is observed in the verbal condition but not in the combined or numeric conditions, where the presence of numeric information is expected to cancel any valence-discrimination relation that might be induced via gain-loss framing.

A final aim of this research was to examine how individual difference measures of cognitive ability and cognitive style predict compliance with the NATO lexicon. In doing so, we expand on previous research showing that numeracy correlates positively with compliance (Wintle et al., 2019). Numeracy refers to an individual's ability to perform basic mathematical operations that would be expected of a data-literate person (e.g., converting a percentage probability into a decimal and knowing that 0.01 is larger than 0.001). Higher levels of numeracy have been shown to facilitate probability assessment and improve the interpretation of numerical data (Lipkus & Peters, 2009). Meanwhile, individuals with low numeracy are shown to rely on non-numerical cues and to be more vulnerable to presentation effects (Reyna et al., 2009). In the present research, in addition to numeracy, we measured differences in verbal reasoning skill and actively open-minded thinking (AOT). Verbal reasoning skill assesses abstract analogical reasoning using language (Bilker et al., 2014), while AOT assesses people's openness to new information and perspectives contrary to their beliefs (Baron et al., 2015). AOT is positively associated with accuracy in probabilistic judgment tasks (Haran et al., 2013; Mellers et al., 2015), and negatively associated with

https://doi.org/10.1017/S1930297500008603 Published online by Cambridge University Press

certain cognitive biases (Baron, 2008; Toplak et al., 2017; West et al., 2008). To the best of our knowledge, verbal reasoning skill and AOT have not been explored in relation to agreement with NBLP schemes for communicating probability. Consistent with Wintle et al. (2019), we hypothesized that numeracy, verbal reasoning ability and AOT would be positively correlated with our agreement measures (Hypothesis 5).

2 Method

2.1 Sampling strategy and participants

Our primary analyses involved factorial analysis of variance (ANOVA) with twelve between-subjects conditions (i.e., Probability format [3] × Probability level [2] × Frame [2]). Using G*Power (Faul et al., 2007), we computed an a priori power analysis for ANOVA with main and interaction effects with η_p^2 = .025, Type I and II error rates set to 5%, df = 2 in the numerator, which returned a sample size of 606. To accommodate ANOVA with an additional nested factor (Table format) we required a sample of 509, half of which overlapped with the sample required for the aforementioned three-way design. Therefore, we estimated a minimum required sample of 866. We oversampled by approximately 40% to offset the chance that we might need to exclude a significant proportion of incoherent responders as we have encountered the need to do in other judgment research (e.g., Mandel, Collins, et al., 2020). A sample of 1,236 participants (52% male) between the ages of 18 and 60 (M = 43.79, SD = 11.77) was recruited using the online crowdsourcing service Qualtrics Panels (https://www.qualtrics.com/). Qualtrics Panels incentivizes participants using a variety of methods that typically correspond to 40%–60% of the per-participant cost charged to researchers. In the present research that corresponds to $6-$9 US, for completion of the full survey set (see procedure and materials). All participants were sampled from Canada or the U.S., and were required to have English as their first language. Participants were prohibited from completing the experiment using a smartphone and were also screened out if they failed a one-item instructional manipulation check designed to test their attention to instructions (Oppenheimer et al., 2009).

2.2 Design

Participants were randomly assigned to 12 conditions in a 3 (Probability format: verbal, combined, numeric) × 2 (Probability level: low, high) × 2 (Frame: positive, negative) between-subjects factorial design. A fourth factor we refer to as table format was manipulated between subjects and nested in the verbal condition. Specifically, participants assigned to the verbal condition were further randomly assigned to either a full-table or partial-table condition. In the full-table condition, participants were shown the full NATO translation table (see Table 1), whereas in the partial-table condition, the numeric equivalents shown in the first column of Table 1 were omitted. In the combined and numeric conditions,

https://doi.org/10.1017/S1930297500008603 Published online by Cambridge University Press

participants were presented with the full table. Probability format refers to whether the intelligence assessment reported in the experimental task stated only the verbal probability term (e.g., *likely*), the verbal term with the numeric range in parentheses (e.g., *likely* [60%–90%]), which we call the combined condition, or only the numeric range (e.g., 60%–90%). Probability level refers to whether the intelligence assessment used a low probability (e.g., *unlikely* [10% –40%]) or a high probability (e.g., *likely* [60%–90%]). Frame refers to whether the outcome was described positively (i.e., half of a group of civilians surviving) or negatively (i.e., half of the group dying).

2.3 Procedure and materials

The experiment was conducted as part of a small set of brief, counterbalanced experiments administered online through Qualtrics. Participants were not informed of the aims of the research until the end of the experiment and they could not alter responses entered on previous screens. At the beginning of the experiment, participants were informed that they would receive information from a hypothetical intelligence report and answer a set of questions. They were introduced to the NATO translation table (partial or full, depending on their condition) and informed that the analyst had used one of the probability terms when making a forecast. Participants were then presented with a hypothetical humanitarian crisis and an intelligence forecast regarding the survival of 1,000 displaced civilians.

After participants reviewed the scenario, the hypothetical intelligence assessment was presented as follows (probability level and frame manipulations shown in brackets):

> Given the current situation on the ground, a senior intelligence analyst specializing in that region assesses
>
> [in the verbal condition] 'It is [likely/unlikely] that half of these civilians will [survive/die].'
>
> [in the combined condition] 'It is [likely (namely, there is a 60% - 90% chance)/unlikely (namely, there is a 10% - 40% chance)] that half of these civilians will [survive/die].'
>
> [in the numeric condition] 'There is a [60% - 90%/10% - 40%] chance that half of these civilians will [survive/die].'

The scenario and intelligence assessment remained visible while participants responded to subsequent questions, whereas the NATO translation table was visible only at the beginning of the experiment. However, before proceeding to the first set of questions, participants had the opportunity to review the NATO translation table (with or without numeric equivalents, depending on their condition) by clicking a clearly labeled button. After proceeding, they were presented with the first set of questions, along with the text of the scenario and intelligence assessment. In the following order, participants were asked to provide their best, lowest, and highest estimates of the probability that the intelligence analyst had in

https://doi.org/10.1017/S1930297500008603 Published online by Cambridge University Press

mind by responding on sliders ranging from 0 to 100 with a default starting position of 0.[2] The three questions were phrased as follows:

> (1) What is your **BEST** estimate of the probability conveyed by the analyst?
>
> (2) and (3) What is the [**LOWEST**, **HIGHEST**] probability the analyst conceivably has in mind?

Participants subsequently completed an additional set of questions, which are the focus of a separate investigation that also includes data from other experiments.[3] After completing the core experimental tasks, participants were given a one-item instructional manipulation check (Oppenheimer et al., 2009). Qualtrics Panels excluded participants who did not answer this task correctly. Participants who correctly answered the instructional manipulation check subsequently completed a 10-item numeracy scale drawing eight questions from Lipkus et al.'s (2001) numeracy scale and two questions from the Berlin Numeracy Test (Cokely et al., 2012); an 8-item verbal skills test comprised verbal analogy questions from the 29-item Penn Verbal Reasoning Test (PVRT; Bilker et al., 2014); and the eight-item actively-open-minded thinking scale from Baron et al. (2015). Finally, participants answered basic demographic questions (i.e., age, sex, and professional experience) to further characterize the sample.

2.4 Agreement measures

We computed three agreement measures. First, in line with earlier studies (Budescu et al., 2009, 2012, 2014; Wintle et al., 2019), we categorized whether best estimates fell within the relevant ranges stipulated by the NATO lexicon and analyzed the proportion of agreeing best estimates (PABE). Our second measure relied on the mean absolute difference (MAD) of the best estimate and the midpoint of the relevant NATO range. This measure reflects the distance between a participant's best estimate and what is arguably the best prototypical point within the relevant stipulated range. However, to enable multivariate analyses with the other agreement measures, we multiplied MAD by -1 so that for all three agreement measures, higher values reflected better agreement. We refer to the negated MAD measure as MAD_{neg}. Our third measure was the mean percentage overlap (MPO) between the participant's range and the stipulated range as shown in Equation 2. In cases where spread was equal to 0 ($n = 78$), PO equaled 100% if the value of the bounds fell within the stipulated range; otherwise, PO equaled 0%.

[2] If a participant wanted to indicate 0 as their response, they would have to move the slider away from 0 and then back.

[3] Briefly, participants were asked to estimate the number of civilians who would [survive/die] by responding on a slider ranging from 0 to 1,000. They also estimated the probability on a 0–100 percent-chance scale that 0 or more civilians will [survive/die]; 100 or more civilians will [survive/die]; 200 or more civilians will [survive/die], and so on up to the probability that all 1,000 civilians will [survive/die].

https://doi.org/10.1017/S1930297500008603 Published online by Cambridge University Press

3 Results

3.1 Preliminary analyses

Thirty-four (2.8%) participants provided lower-bound estimates that exceeded their upper-bound estimates. These cases were removed (revised N = 1,202). Approximately 19% of remaining participants provided best estimates that fell outside the credible interval defined by their lower- and upper-bound estimates. These violations were independent of probability format, probability level, or frame based on chi-square tests (all $p > .28$). Wintle et al. (2019) rearranged such estimates into their logical order. However, we neither altered nor removed them.

3.2 Primary measures of equivalence

3.2.1 Spread

In the verbal condition, the effect of table format on spread (i.e., the upper bound minus the lower bound) was not statistically significant ($t[599] = 0.40$, $p = .69$, Cohen's $d = 0.03$). Therefore, we collapsed over this nested factor in analyses of other effects, and we used the full sample. As an exploratory analysis, we conducted a three-way (Probability format × Probability level × Frame) ANOVA on spread. None of the main or interaction effects were significant (all $p >.07$).[4] The grand mean of spread was 31.91 [30.71, 33.10].

3.2.2 Best estimates

In the verbal condition, the effect of table format on best estimates was not statistically significant ($t[599] = 0.75$, $p = .45$, Cohen's $d = 0.06$). Therefore, this nested factor was collapsed over analyses of other effects. We conducted a three-way (Probability format × Probability level × Frame) between-subjects factorial ANOVA on best estimates. As expected, the main effect of probability level was significant ($F[1, 1190] = 470.95$, $p < .001$, $\eta_p^2 = .284$). The mean estimate of the low probability was 41.48 [39.93, 43.02] and the mean estimate of the high probability was 65.52 [63.99, 67.05].[5] However, probability level significantly interacted with probability format ($F[2, 1190] = 34.33$, $p < .001$, $\eta_p^2 = .055$). Figure 1 plots the interaction effect, which shows that the discrimination between the low and high probabilities is significantly lower in the verbal condition than in the combined or numeric conditions, the latter two of which are virtually indistinguishable. No other effect in the model was statistically significant (all $p > .5$). Therefore, in the present experiment, we rejected Hypothesis 4 and find no evidence that positive/negative framing of outcomes affects the discrimination between low and high probability terms.

[4]This and all subsequent ANOVAs use Type III sum of squares. The only effect of marginal significance was the three-way interaction, which we neither predicted nor do we wish to speculate about.

[5]We report 95% confidence intervals in square brackets following the relevant estimate.

https://doi.org/10.1017/S1930297500008603 Published online by Cambridge University Press

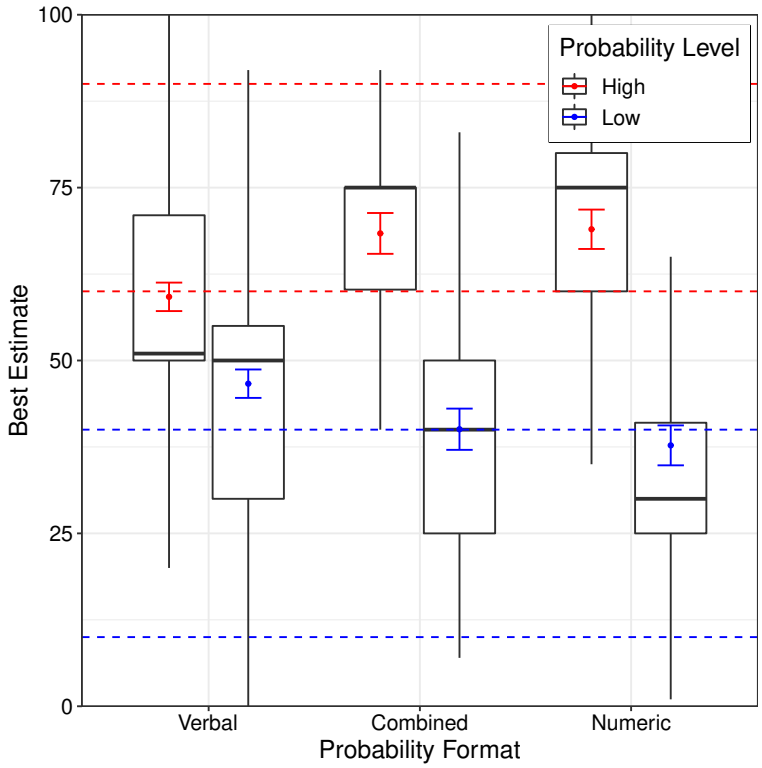

FIGURE 1: Summary of best estimates by probability format and probability level. Dashed lines represent NATO numeric range equivalents for the verbal probability terms *likely* and *unlikely*, respectively. Box-and-whisker plots are from sample data, whereas the error bars are marginal means and 95% confidence intervals from the ANOVA model.

As Figure 1 shows, the median probabilities for the terms *unlikely* and *likely* are virtually indistinguishable and they fall on or very close to 50%, in contrast to the medians observed in the combined and numeric conditions. This raises the possibility that (despite the initial starting position of 0 on the slider scale) a significant proportion of participants may have responded with 50% as their best estimate to reflect a "don't know" response, thus producing a *fifty-fifty blip* (Bruine de Bruin et al., 2002; Fischhoff & Bruine de Bruin, 1999). More specifically, the results in Figure 1 suggest that the proportion of fifty-fifty responders is significantly greater in the verbal condition than in the combined or numeric conditions. We tested this hypothesis using strict and loose classification methods. For the strict method, we dummy coded participants whose best estimates equaled 50% as 1 and otherwise as 0. For the loose method, we coded responses between 49% and 51% inclusive as 1 and otherwise as 0. The loose method reflects the fact that the slider was quite sensitive to movement and someone intending to respond with 50% might easily end up a point higher or lower on the scale. As Table 2 shows, the percentage of fifty-fifty responders was significantly greater in the verbal condition than in the combined or numeric conditions. This was the case for both the strict and loose methods. As well, using the strict method, the percentage of fifty-fifty responders was marginally greater in the combined condition than in the numeric condition.

https://doi.org/10.1017/S1930297500008603 Published online by Cambridge University Press

Table 2: Percentage of fifty-fifty responders by probability format and method.

	Probability format			Verb.-Comb.		Verb.-Num.		Comb.-Num.	
Method	Verb.	Comb.	Num.	Z	p	Z	p	Z	p
Strict	23.3	5.5	2.6	−6.11	.000	−7.67	.000	−1.84	.066
Loose	37.2	9.0	5.8	−8.11	.000	−9.49	.000	−1.49	.135

Note. Verb., Comb., and Num. stand for verbal, combined and numeric conditions, respectively. Pairwise comparisons are based on Mann-Whitney U test. Significance values are two-tailed.

The preceding analyses naturally raise the question of whether the discrimination between low and high probabilities is still affected by probability format if fifty-fifty responders are excluded, as the exclusion of this subset must attenuate the interaction effect plotted in Figure 1. Accordingly, we recomputed the three-way ANOVA on best estimates after excluding those who met the definition of fifty-fifty responders by the loose criterion in order to retest the probability level × probability format interaction effect. As expected, the two-way interaction effect was attenuated and only approached conventional significance levels ($F[2, 734] = 2.37$, $p = .094$, $\eta_p^2 = .006$). Figure 2 plots this interaction effect. Compared to Figure 1, estimates in the verbal condition are much less regressive. The median probabilities assigned to the terms *unlikely* and *likely* now fall in the stipulated ranges, although the mean of the term *unlikely* still falls outside the stipulated range.

To provide a direct test of the hypothesis that participants' best estimates were more regressive in the verbal condition than in the combined or numeric conditions, even after excluding fifty-fifty responders, we computed two extremity scores, E_P and E_L, as follows:

$$E_p = 50 - B_e \text{ if} PL = \text{low} \ \wedge \ E_p = B_e - 50 \text{ if} PL = \text{high} \tag{3}$$

$$E_L = \max(50 - B_e, 0) \text{ if} PL = \text{low} \ \wedge \ E_p = \max(B_e - 50, 0) \text{ if} PL = \text{high}, \tag{4}$$

where B_e is the participant's best estimate and PL stands for the design factor, probability level. The subscripts P and L on E stand for *punitive* and *lenient*, respectively. The higher the value of E, the more extreme (or less regressive) the participant's best estimate is provided it is correctly located relative to 50% — namely, provided best estimates for low probabilities are not more than 50% and best estimates for high probabilities are not less than 50%. Violations of these constraints yield negative "anti-extremity" values for E_P and values of 0 for E_L. Both measures differ, therefore, from one that merely scores the absolute difference between 50 and B_e. The absolute difference would, of course, fail to differentiate a participant who indicates that unlikely means 20% from one who indicates that it means 80%, treating normative and perverse forms of extremity at a constant magnitude equally.

After excluding fifty-fifty responders using the loose criterion, a one-way (Probability format) ANOVA computed on punitive extremity, E_P, was statistically significant ($F[2, 743]$

https://doi.org/10.1017/S1930297500008603 Published online by Cambridge University Press

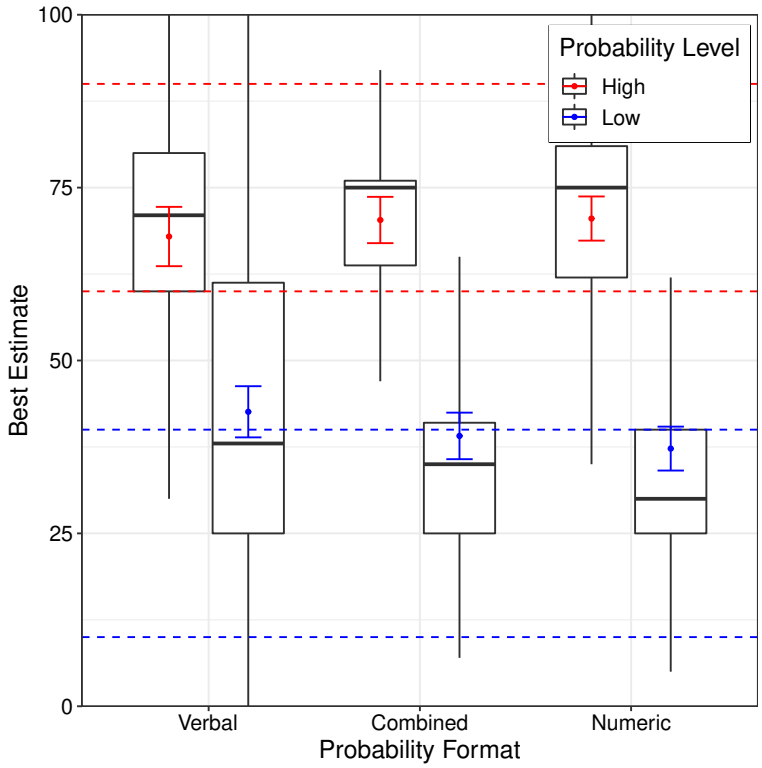

FIGURE 2: Summary of best estimates by probability format and probability level, excluding fifty-fifty responders. Dashed lines represent NATO numeric range equivalents for the verbal probability terms *likely* and *unlikely*, respectively. Box-and-whisker plots are from sample data, whereas the error bars are marginal means and 95% confidence intervals from the ANOVA model.

= 3.05, $p = .048$, $\eta_p^2 = .008$). Compared to participants' best estimates in the verbal condition (M = 12.05 [9.18, 14.92]), those in the numeric condition (M = 16.60 [14.30, 18.91]) were significantly more extreme ($p = .041$ by Tukey's HSD test), whereas those in the combined condition (M = 15.53 [13.10, 17.96]) did not significantly differ from those in either the verbal condition ($p = .17$) or the numeric condition ($p = .81$). We conducted a second test that was identical to the preceding test, except that we swapped the punitive measure for the lenient measure of extremity, E_L. The main effect, once again, was statistically significant (F[2, 743] = 3.32, $p = .037$, $\eta_p^2 = .009$). Compared to participants' best estimates in the verbal condition (M = 17.05 [15.50, 18.90]), those in the numeric condition (M = 20.15 [18.66, 21.63]) were significantly more extreme ($p = .029$ by Tukey's HSD test), whereas those in the combined condition (M = 18.67 [17.11, 20.24]) did not significantly differ from those in either the verbal condition ($p = .39$) or the numeric condition ($p = .37$). Therefore, even after removing fifty-fifty responders, participants' best estimates were more regressive when they were presented with verbal probabilities than when they were presented with numeric ranges.

Our final analyses in this section aim to shed light on the causal bases for the observed

https://doi.org/10.1017/S1930297500008603 Published online by Cambridge University Press

fifty-fifty blip, which was most strongly manifested in the verbal condition. Bruine de Bruin et al. (2000) found that the fifty-fifty blip was stronger for less numerate and younger participants (i.e., youth vs. adults), and for events that were singular rather than distributional and, therefore, more likely to be associated with epistemic rather than aleatory uncertainty. We tested support for a similar pattern of results in the present research. However, numeracy did not significantly differ between participants who gave fifty-fifty responses and those who did not (t[900] = 1.16, p = .25). In our study, age did differ, but fifty-fifty responders, on average, were *older* (M = 46.97, SD = 10.48) than the remainder of the sample (M = 43.16, SD = 11.79; t[244.40] = 4.04, p < .001).[6] Moreover, the significance of these effects was virtually unchanged if the sample is restricted to participants in the verbal condition.

Another possibility suggested by past research (e.g., Bruine de Bruin et al., 2000) is that fifty-fifty responders are less epistemically certain than their counterparts about their estimate. If so, we might expect the spread of the lower and upper bounds of their credible intervals to be greater for fifty-fifty responders than their counterparts, as wider spreads represent greater uncertainty. To the contrary, spread was significantly greater in the subsample that did not respond fifty-fifty (M = 32.88, SD = 16.97) than among fifty-fifty responders (M = 26.53, SD = 21.81; t[196.09] = 3.43, p = .001). The variances of these subsamples were significantly different too, according to Levene's test (F = 30.54, p < .001). The smaller variance among fifty-fifty responders suggests an alternative hypothesis: perhaps there is a corresponding *zero-spread blip* for credible intervals among fifty-fifty responders, consistent with the hypothesis that, for some individuals, verbal probabilities simply are not interpreted in quantitative terms. In fact, if we examine the distribution of spreads in the verbal condition, we find that there is a zero-spread blip for fifty-fifty responders, whereas there is no such blip for the counterpart subsample. Whereas only 3.7% of the latter subsample gave bounds that yielded a spread of 0, fully 25% of fifty-fifty responders did so. In fact, the zero-spread blip represented the mode in that subsample.[7] The difference in these percentages is highly significant (by Mann-Whitney U test, z = -10.53, p < .001) and large: fifty-fifty responders in the verbal condition are 6.8 times more likely to indicate a zero spread than their "non-50%" counterparts in the verbal condition.

Finally, if we compare the percentage of participants who gave fifty-fifty responses *and* had zero-spread (using the loose criterion in both cases), we find 11.0% in the verbal condition, whereas the percentage is 0.3% in each of the other two conditions — namely, participants were 36.7 times more likely to exhibit this pattern of response in the verbal condition than in the combined or numeric conditions. These findings suggest that just over 10% of people asked to interpret the numeric meaning of verbal probability show signs of what we call *representational mapping incapacity*. For these individuals, it may be difficult to conceive of a mapping from verbal to numeric probabilities. If so, this difficulty does not appear to be related to numeracy, as the minority exhibiting this pattern in the verbal

[6]We use the loose criterion for classifying fifty-fifty responders in this and subsequent analyses.

[7]If we allow for a loose scoring of spread (i.e., ≤ 2), then the percentage is 29.5% in the fifty-fifty subsample and 5.8% in the remaining subsample.

https://doi.org/10.1017/S1930297500008603 Published online by Cambridge University Press

condition did not significantly differ from the majority who did not exhibit the "50±0" pattern (t[299] = 0.86, p = .39).

3.3 Agreement

3.3.1 Individual differences in cognitive performance and style

We first examined whether the three agreement measures (i.e., MPO, MAD_{neg}, and PABE) were correlated with numeracy, PVRT, and AOT. Consistent with earlier findings (Wintle et al., 2019), as Table 3 shows, greater agreement (across all three measures) was positively related to higher numeracy, verbal-reasoning skill, and actively open-minded thinking. Therefore, we found consistent support across multiple tests of Hypothesis 5. None of these individual difference measures significantly differed across any of our experimental manipulations.

TABLE 3: Pearson correlation matrix.

	MAD_{neg}	PABE	Numeracy	PVRT	AOT
(1) MPO	.58**	.66**	.20**	.18**	.18**
(2) MAD_{neg}		.78**	.24**	.15**	.14**
(3) PABE			.18**	.13**	.15**
(4) Numeracy				.45**	.36**
(5) PVRT					.37**

*p < .05, **p < .01.

3.3.2 Table format

Recall that we predicted that agreement would be better in the full table condition than in the partial table condition (Hypothesis 1). To examine the effect of table format on agreement we restricted our analyses to the subset of cases in the verbal condition where format was varied (n = 601) and conducted a one-way (Table format) multivariate ANOVA (MANOVA) on the three agreement measures. The multivariate effect of table format was not statistically significant (F[3, 597] = 1.82, p = .143, η_p^2 = .009). However, the univariate results were mixed. Both agreement measures that relied on best estimates (i.e., PABE and MAD_{neg}) were not statistically significant (both p > .115), whereas the measure that relied on lower and upper bounds (i.e., MPO) was significant (F[1, 599] = 5.24, p = .020, η_p^2 = .009). Using the MPO measure, agreement was, in fact, better in the full table condition (M = 0.40 [0.36, 0.44]) than in the partial table condition (M = 0.33 [0.29, 0.37]). These findings therefore provide partial support for Hypothesis 1. Evidently, providing numeric ranges for stipulated terms helps foster agreement, but only on measures that rely on range input for calculating agreement.

https://doi.org/10.1017/S1930297500008603 Published online by Cambridge University Press

3.3.3 Probability format, probability level, and frame

Turning to cases presented with the full translation table (n = 902), we examined the three agreement measures in a three-way (Probability format × Probability level × Frame) factorial MANOVA. There was a significant multivariate main effect of probability format (F[6, 1778] = 27.29, p < .001, η_p^2 = .084). All three univariate F tests were significant at p < .001. Table 4 shows that for each of the three agreement measures, agreement in the verbal condition was significantly poorer than in the combined and numeric conditions, and the latter two conditions did not significantly differ. Moreover, the effect of probability format did not significantly interact with probability level or frame (smallest p = .312). These findings strongly support Hypotheses 2 and 3.

The multivariate main effect of probability level was also statistically significant (F[3, 888] = 23.48, p < .001, η_p^2 = .073). However, the results of the univariate F tests were at odds. Using MPO, agreement was better for the lower probability (M = 0.67 [0.64, 0.71]) than for the higher probability (M = 0.57 [0.53, 0.60]; F[1, 890] = 16.89, p < .001, η_p^2 = .019). For MAD_{neg}, the effect was in the opposite direction, with worse agreement for the lower probability (M = −18.34 [−19.78, −16.91]) than for the higher probability (M = −14.79 [−16.22, −13.36], F[1, 890] = 11.83, p = .001, η_p^2 = .013). Finally, for PABE, the effect was not significant (F[1, 890] = 0.12, p = .731, η_p^2 = .000). No other effect in the MANOVA model was statistically significant at α = .05.

Finally, we recomputed the MANOVA on agreement measures with fifty-fifty responders excluded based on the loose criterion. The new model yielded the same significant effects: for probability format (multivariate F[6, 1466] = 8.18, p < .001, η_p^2 = .032), for probability level (multivariate F[3, 732] = 17.37, p < .001, η_p^2 = .066). Therefore, the findings are robust regardless of whether fifty-fifty responders are included or excluded from the analysis.

3.3.4 Decision to review the NBLP scheme

Recall that participants were given the option of reviewing NATO's NBLP scheme prior to giving their probability estimates. We examined whether the effect of probability format reported earlier interacted with participants' decision to either review the table or not. Among participants in the full table condition, 492 (54.5%) reviewed the table before providing their probability equivalent (dummy coded as 1 and otherwise as 0). The effect of probability format on this percentage only approached statistical significance (χ^2[2, N = 902] = 4.86, p = .088); percentages who chose to review equal 52.8%, 51.0%, and 59.5% in the verbal, combined and numeric conditions, respectively.

We conducted a two-way (Probability Format × Review) MANOVA on the agreement measures. In particular, we sought to examine whether there was a significant interaction effect. Perhaps the poor agreement in the verbal condition compared to the combined and numeric conditions was due to participants' failure to attend to the NBLP scheme. If so, we should observe a stronger simple effect of review in the verbal condition than in the other two

https://doi.org/10.1017/S1930297500008603 Published online by Cambridge University Press

Table 4: Agreement measures by probability format

		95% confidence interval bounds	
Probability format	MPO	Lower	Upper
Verbal	39.75^{a}	35.49	44.00
Combined	71.76^{b}	67.41	76.10
Numeric	74.63^{b}	70.44	78.83
	MAD_{neg}		
Verbal	-20.58_{c}	−22.34	−18.83
Combined	-14.12_{d}	−15.91	−12.34
Numeric	-15.00_{d}	−16.72	−13.27
	PABE		
Verbal	39.88^{e}	34.53	45.23
Combined	69.02^{f}	63.56	74.48
Numeric	69.18^{f}	63.90	74.45

Note. Values in the second column that do not share the same superscript within measure significantly differ at $p < .001$ by Tukey's HSD test and those sharing a subscript do not significantly differ at $\alpha = .05$.

conditions. First, we observed a multivariate main effect of review ($F[3, 894] = 8.08$, $p < .001$, $\eta_p^2 = .026$).[8] However, only the univariate F test on PABE was statistically significant ($F[1, 896] = 7.96$, $p = .005$, $\eta_p^2 = .009$). In this case, participants who reviewed the scheme prior to providing numeric equivalents showed better agreement ($M = 0.64$ [0.60, 0.68]) than those who did not review the scheme ($M = 0.55$ [0.50, 0.59]). More importantly, the multivariate interaction effect was statistically significant ($F[6, 1790] = 2.60$, $p = .016$, $\eta_p^2 = .009$). Moreover, all univariate F tests for the interaction effect were significant at $\alpha = .01$.

To simplify the presentation of the interaction across agreement measures, we standardized the three agreement measures and averaged them to form an agreement scale, which had good reliability, Cronbach's $\alpha = .86$. Figure 3 plots the interaction effect. As anticipated, reviewing the scheme improved agreement in the verbal condition. The simple effect on the composite measure was statistically significant ($F[1, 299] = 11.20$, $p = .001$, $\eta_p^2 = .036$). In contrast, in the combined condition, the decision to review the scheme had no significant effect ($F[1, 288] = 0.37$, $p = .545$, $\eta_p^2 = .001$). Finally, in the numeric condition, there was a marginally significant effect in the opposite direction to that observed in the verbal condition ($F[1, 309] = 2.14$, $p = .081$, $\eta_p{}^2 = .010$). That is, participants who

[8]The main effect of probability format was significant as well, but this was already reported in the previous MANOVA.

https://doi.org/10.1017/S1930297500008603 Published online by Cambridge University Press

did not choose to review the scheme showed better agreement than those who chose to review it. It is also evident from Figure 3 that the simple effect of presentation format was significant. In particular, it is clear from the non-overlapping 95% confidence intervals that, even among those participants that reviewed the scheme immediately prior to judging the numeric equivalents, agreement in the verbal condition is surpassed by that in the combined and numeric conditions.

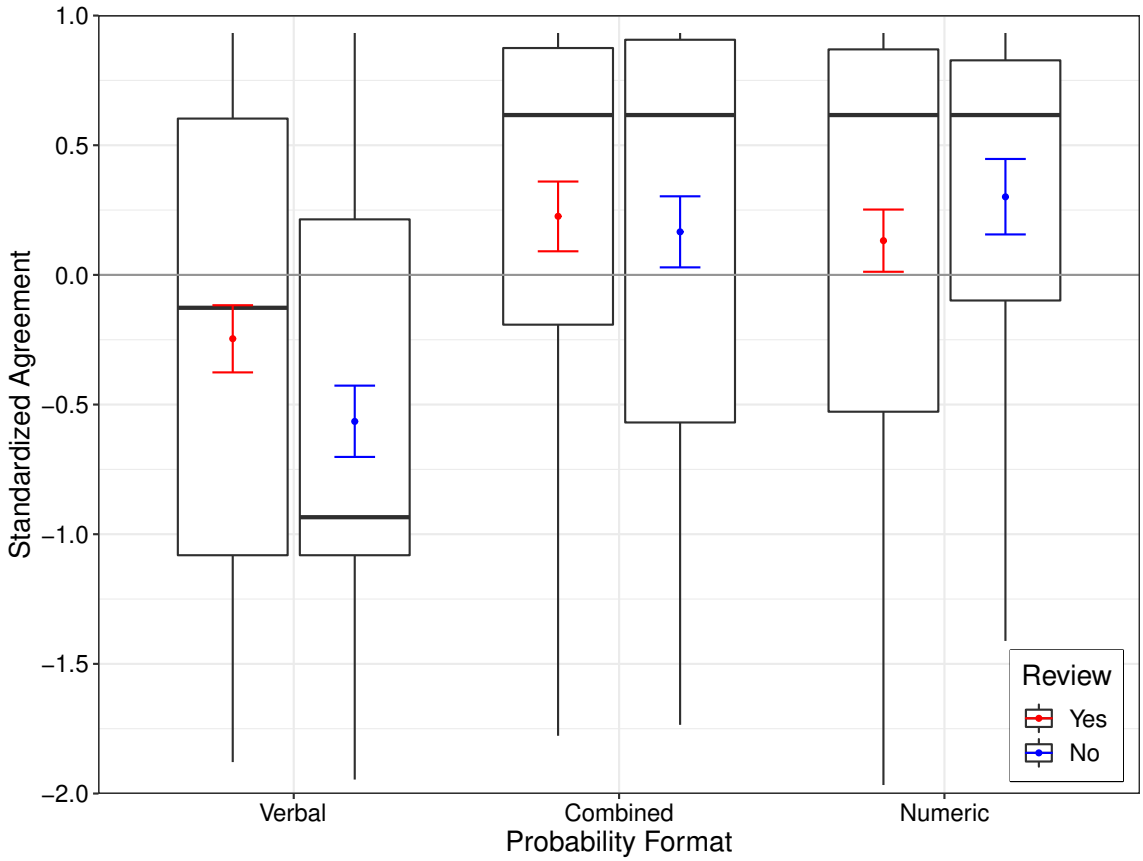

FIGURE 3: Standardized agreement by probability format and review.

3.3.5 Extremity, presentation format, and agreement

The preceding results showed that the extremity of best estimates remained affected by probability format after excluding fifty-fifty responders. As well, after fifty-fifty responders were excluded, presentation format continued to significantly affect agreement. Taken together, these findings suggest that the effect of probability format on agreement is mediated at least partly by extremity. In this final analysis, we tested this hypothesis directly. Figure 4 shows the standardized regression weights for links in the model in which extremity (using E_P) mediates the effect of probability format on agreement (using the composite measure). The attenuation of the probability format effect on agreement after controlling for the mediator was statistically significant (Sobel test $z = 5.19$, $p < .001$). Even after

https://doi.org/10.1017/S1930297500008603 Published online by Cambridge University Press

controlling for extremity, the predictive effect of probability format was still significant. These results suggest that extremity partially mediates the effect of probability format on agreement (Baron & Kenny, 1986).

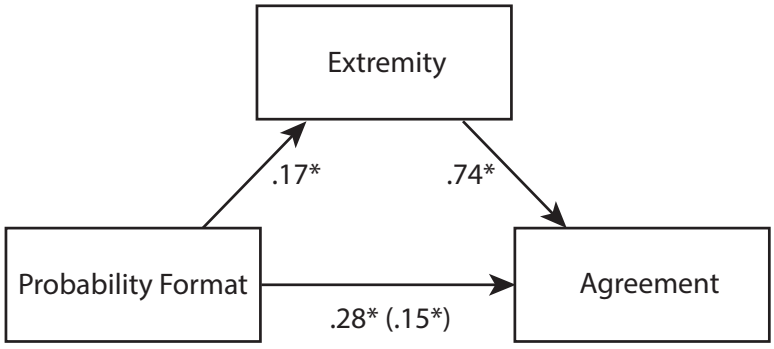

FIGURE 4: Mediator model of probability format effect on agreement. *$p < .001$.

4 Discussion

This research tested several hypotheses about contemporary organizational approaches to communicating probabilities to end-users, which rely on NBLP schemes. We investigated agreement with one such scheme used by NATO in the context of military intelligence production and dissemination and using verbal probability terms (i.e., *likely* and *unlikely*) that are widely employed in other NBLP schemes (for other examples, see Dhami & Mandel, 2020; Ho et al., 2015; Morgan, 1998). One issue we sought to address was whether the numerically-bounded component of NBLP schemes conferred an advantage in terms of fostering agreement. That is, do schemes that provide numeric ranges as semantic anchors fare better than comparable schemes that provide only an ordered set of verbal probabilities? Across the three measures of agreement we computed, only one (based on the proportion of overlap) showed a benefit to using ranges in the scheme. It is noteworthy that the one measure of agreement that was improved by providing ranges was itself range-dependent, having been computed using lower and upper bound values. In contrast, the other measures relied on the participant's best estimate of the numeric meaning of the relevant term. Therefore, these findings call into question the effectiveness of attempting to stipulate the meaning of verbal probabilities by assigning numeric ranges to them, as organizations have been prone to do and as some researchers have recommended as a comprise in light of the recalcitrant attitude organizations exhibit towards the use of numeric probabilities (Beyth-Marom, 1982). Instead, the findings lend support to recent proposals recommending that organizations and professional bodies communicate probabilities to end-users with numeric probability ranges that can be expressed with more or less precision, as required (e.g., Dhami & Mandel, 2020; European Food Safety Authority et al., 2018; Friedman, 2019; Mandel & Irwin, 2020).

https://doi.org/10.1017/S1930297500008603 Published online by Cambridge University Press

The present research also extended previous work on users' agreement with NBLP schemes by examining a purely numeric condition in which only ranges (corresponding to those depicted in NATO scheme) were used in assessments. Consistent with earlier studies (Budescu et al., 2009, 2012, 2014; Wintle et al., 2019), we found that the combined format produced significantly better agreement than the verbal format. A novel result, however, was that agreement using the purely numeric format was just as good as the combined format on all agreement measures; that is, regardless of whether agreement was calculated on the basis of participants' best estimates or their lower and upper bounds. Simply put, there was a substantial cost imposed on agreement if numeric range information was not included in the assessment, yet there was no observable cost to agreement if verbal probability information was omitted. This pattern was evident even if we examined only the subsample of participants who took care to review the NATO standard immediately before making their judgments. Taken together, these findings show that not only do numeric probabilities improve upon the communicative function of verbal probabilities when they are embedded directly into probabilistic statements, as some have noted (e.g., Budescu et al., 2014; Ho et al., 2015; Patt & Dessai, 2005), but critically, numeric probabilities can replace the use of verbal probabilities insofar as fostering communicative agreement about degrees of probability is the main goal of communication.

The results further show that the cost imposed on agreement by using verbal probabilities is associated with a lack of discrimination between low and high probability terms. Despite having just read the NATO scheme moments before making their judgments, participants, on average, provided best estimates that were too high for the term *unlikely* and too low for the term *likely*. In fact, their estimates were so regressive that their median probabilities were virtually indistinguishable and centered on 50%. As we observed, the regression toward the midpoint of the probability scale was, in part, due to the fact that there were significantly more fifty-fifty responders in the verbal condition than in the combined or numeric conditions. However, even after removing the subsample of fifty-fifty responders, best estimates were still significantly more regressive in the verbal condition than in the numeric condition, and agreement was still lower in the verbal condition than in the combined or numeric conditions. In fact, the extremity of participants' best estimates partially mediated the probability format effect on agreement.

The preceding findings suggest that the use of verbal probabilities to communicate probability levels can undermine information value to end users in two distinct ways. First, by increasing the proportion of fifty-fifty responses, verbal probabilities increase ambiguity about the meaning of the terms. This is because 50% can represent a first-order probability judgment or it could represent the sender's utter epistemic uncertainty about what probability to assign (Bruine de Bruin et al., 2002; Fischhoff & Bruin de Bruin, 1999). In the present research, participants in the verbal condition were 6.4 times (using the loose criterion) to 9.0 times (using the strict criterion) more likely to give a fifty-fifty response than participants in the numeric condition. This effect of probability format on fifty-fifty responses represents

https://doi.org/10.1017/S1930297500008603 Published online by Cambridge University Press

a large increase in ambiguity production. Second, by making probability judgments appear less extreme to receivers, verbal formats for communicating uncertainty are likely to water down the information value to end-users. Since the value of probabilistic assessments is judged to be a function of informativeness and accuracy (Yaniv & Foster, 1995), the regressiveness of verbal probabilities is likely to discount the value of such assessments to end-users, perhaps partly explaining the communication mode preference paradox noted earlier (e.g., Erev & Cohen, 1990).

While requiring further research, the present findings shed light on the causal bases of these effects. Contrary to Bruine de Bruin et al. (2000), we did not find that fifty-fifty responders were less numerate or younger than those who did not exhibit that response. In fact, we found that older participants were more likely to exhibit the fifty-fifty blip. The comparison of age effects across these studies, however, must be interpreted cautiously since Bruine de Bruin et al. (2000) compared youth and adults, whereas we examined age within an adult sample. Bruine de Bruin et al. (2000) also found that the fifty-fifty blip was associated with greater epistemic uncertainty, which we hypothesized might manifest in the present experiment as greater spread. We found, however, the opposite result: spread was greater for participants who did not give a fifty-fifty response.

This last result, however, suggested an alternative hypothesis that garnered support. That is, we reasoned that a nontrivial proportion of fifty-fifty responders in the verbal condition might simply fail to interpret verbal probabilities in quantitative terms and also produce a zero spread. In support of this hypothesis, in the verbal condition, fifty-fifty responders were about 7 times more likely to indicate a zero spread than their "non-50%" counterparts. The comparison across conditions was even more striking, with participants yielding the "50±0" pattern about 37 times more frequently in the verbal condition than in combined or numeric conditions. As noted earlier, the "50±0" pattern we observed in just over 10% of participants in the verbal condition suggests that these individuals have a *representational mapping incapacity*. For these individuals, it may be difficult to conceive of a mapping from verbal to numeric probabilities — a difficulty unrelated to numeracy. Such results are consistent with Mandel et al. (2021) which found that, whereas numeracy was related to the accuracy and coherence of arithmetic computations of averages and products among participants asked to compute these results with numeric probabilities, numeracy was not correlated with these performance measures among participants who received verbal probabilities as inputs to computation. Mandel et al. (2021) suggested that the findings reveal a *differential schematicity effect* in which the schema for arithmetic computing is less available when individuals are given verbal probabilities rather than numeric probabilities to work with. These authors also found that the mapping of verbal probabilities to numeric equivalents was unreliable even though such mappings were elicited in a brief timespan and the task context did not vary between mappings. Taken together, such findings indicate that some individuals cannot map verbal probabilities to numeric probabilities, even if allowances for imprecision are permitted (as in the present research).

https://doi.org/10.1017/S1930297500008603 Published online by Cambridge University Press

In terms of the what we have until now called the "regressiveness" of best estimates, which was found to be greater in the verbal condition than in the numeric condition even after removing fifty-fifty responders from the sample, it is perhaps more accurate to describe this as "response contraction bias", although similar response tendencies have been called regression effects (Stevens & Greenbaum, 1966). As Poulton (1994) explains, true regression effects are caused by variability, whereas response contraction is due to the effect that the central value of a scale, serving as a psychological default, has on estimation through an anchoring and adjustment process. Much earlier, Hollingworth (1910) referred to this as the *central tendency of judgment*. This strikes us as applicable in the present context since the midpoint of the probability scale, in fact, has certain default properties. It is the expected value of random probability draws and it corresponds to the point of maximum uncertainty when the possibility space is binary. This default is often a valid starting point when orienting to a new stimulus that may be present or absent, or to a new hypothesis that may be true or false. As a corrective for response contraction bias, Poulton (1994) recommends using the extreme values of the scale as anchors. In the present experiment, we used 0 as the default. If Poulton (1994) is correct, we might have anticipated even greater response contraction in the verbal condition if the default had been set on 50%; a test that could be performed in future research.

The present research did not show the "valence effect" shown in a few other studies (Dhami & Mandel, 2021; Mandel, 2015a; Mullet & Rivet, 1991). Unlike the earlier studies, which manipulated the events such that one was in some way a *reflection* of the other (e.g., "success vs. failure" or "fit vs. not fit"), in the present research the same event was *framed* either in terms of lives to be saved or lives to be lost. Whereas manipulations of reflection refer to different events, manipulations of frame refer to the same events that are "merely" described differently. It is possible that this difference accounts for the failure to get the result. However, it is also possible that the valence effect is not particularly robust. Since each of the earlier studies used a distinct task structure, it is premature to judge whether the comparative difference between this research and the earlier studies may be attributable to the framing context. In fact, it is possible that the earlier findings are not themselves reflective of a unitary valence effect. For instance, in Mandel (2015a) success versus failure was used to manipulate valence, whereas affirmative versus negational statements were used in Dhami and Mandel (2021). Boundary conditions for valence effects on the interpretation of verbal probabilities could be explored in future research. Our findings do, however, add to at least one other study showing no interaction of probability formats and frames (Liu et al., 2020). Clearly, this is an area that is ripe for future research.

4.1 Policy Implications

Taken together, the findings of this research call into question current practices that use NBLP schemes, such as those used in climate science communication (e.g., Lewis et al., 2019; Mastrandrea et al., 2011), national security intelligence (Office of the Director of

https://doi.org/10.1017/S1930297500008603 Published online by Cambridge University Press

National Intelligence, 2015; NATO, 2016), and other organizations (e.g., Morgan, 1998). Our findings add to those of other studies (e.g., Budescu et al., 2009, 2012, 2014; Ho et al., 2015; Wintle et al., 2019) that suggest that NBLP schemes are unlikely to achieve their goal of ensuring a high degree of agreement between senders and receivers of uncertain estimates. The present findings show that such schemes do not even ensure that probability terms with different directionality such as *unlikely* and *likely* deflect in opposite directions from fifty-fifty.

Earlier studies (Budescu et al., 2014; Ho et al., 2015; Patt & Dessai, 2005, Wintle et al., 2019) have identified limitations of these schemes. For instance, noting that most schemes, with some noteworthy exceptions (e.g., Barnes, 2016), are formulated by BOGSAT (i.e., "bunch of guys/gals sitting around the table"; e.g., Marvin, 2020), and illustrating how more general calls for the application of scientific research to methodological problems in domains such as intelligence analysis could be conducted (e.g., Chang et al., 2018; Dhami et al., 2015). Ho et al., (2015) aimed to show that the setting of numeric ranges on probability terms could be determined by empirical data and model fitting with a resulting increase in agreement compared to existing NBLP schemes. The recommendations given in earlier work, as noted earlier, have also focused on trying to repair deficiencies by embedding the probability ranges not only into the lexicons used by organizations but into each statement that uses probability terms from the relevant lexicon, as captured in the combined format. Given that (a) we found agreement to be as good using numeric ranges alone as using the combined format, (b) the tendency toward regressiveness in the combined condition fell between the verbal and numeric conditions, and (c) Knapp et al. (2016) found that risk assessments were more realistic following information in a numeric format than in a combined format, we question the utility of imposing NBLP schemes on senders and receivers. As well, the use of numeric ranges in specific assessments to clarify the meaning of vague probability terms runs the risk of being misinterpreted as credible intervals on the probability of events referenced in the substantive assessments, yet this is not what the ranges are intended to signify (Mandel & Irwin, 2020).

Instead, our findings support recommendations for organizations to use numeric probabilities either as point values (with or without margins of error) or as numeric ranges without the use of linguistic probabilities in their communications (Dhami & Mandel, 2020; Friedman, 2019; Mandel, Wallsten et al., 2021). If numeric ranges were unshackled from vague verbal probabilities, they could, in fact, be used as credible intervals on the probability of focal events referenced in substantive assessments and there would be no risk of confusing intervals meant to define terms with intervals that are issue-specific. This would provide decision-makers with useful information both about the probability of events and the uncertainty of the assessment (i.e., depending on the spread).

Of course, numeric quantifiers (i.e., the use of numeric values, precise or imprecise, in linguistic contexts) can still be ambiguous. It is not always obvious whether numeric quantifiers refer to exact values ("precisely p"), lower bounds ("at least p"), upper bounds

https://doi.org/10.1017/S1930297500008603 Published online by Cambridge University Press

("at most *p*"), or fuzzy numbers ("roughly *p*") (e.g., Geurts & Nouwen, 2007; Mandel, 2014). Numeric ranges may also be interpreted in a variable manner, such that different end-users may draw quite different conclusions about the underlying probability distributions and such conclusions tend to be biased by end-users' worldviews in a belief-congruent manner (Dieckmann et al., 2017), although the use of best estimates along with upper and lower bounds can serve to reduce such variability (Dieckmann et al., 2015). Finally, there is little doubt that a principal reason senders prefer verbal probabilities to their numeric alternatives is that they are easier and "more natural" to produce (Wallsten et al., 1993). In contexts where ease is a concern that outweighs transparency, we cannot recommend against the use of verbal probabilities. Moreover, if ease is a significant concern, then the use of NBLP schemes may be preferable to a no-scheme alternative since it could, at least, help steer senders away from the vaguest expressions to which they may be inclined. For instance, a study of probability expressions used in oral radiology found that the expressions used most frequently tended to have the widest range of meanings (Stheeman et al., 1993). Yet if especially vague terms, such as *realistic possibility*, which was recently used in the UK intelligence community's NBLP scheme (Dhami & Mandel, 2020), are selected, such schemes could institutionalize rather than avoid the worst possible choices of terminology.

We further note that justifications for the use of NBLP schemes sometimes turn on the view that receivers do not have the requisite numeracy skills to correctly process numeric probability information (e.g., Lewis et al., 2019). Our findings, however, suggest that low numeracy skill, along with lower verbal reasoning ability and an actively open-minded thinking disposition, also portend difficulty with the proper application of NBLP schemes. In the present research, each measure of agreement was directly related to numeracy and these other measures, calling into question how well they serve the information interests of individuals with lower numeracy. This conclusion suggests that efforts might be better focused on numeracy education. Such education could focus on how to update probabilistic beliefs more coherently (Mandel, 2015b) and use comparison classes (Chang et al., 2016) as well as on overcoming popular misconceptions about quantifying uncertainty, such as the view that assigning numbers to probabilities implies they are scientific estimates (Mandel & Irwin, 2020).

To sum up, in terms of vagueness, numeric probabilities pale in comparison to verbal probabilities. The idea that such vagueness can be brushed away by providing NBLP schemes that stipulate the semantic meaning of probability phrases has been attractive, if not outright seductive, to many organizations tasked with delivering uncertain estimates to diverse audiences. Unfortunately, over multiple studies including the present research, the same notion has garnered virtually no empirical support. NBLP schemes might seem to be a good solution, but a growing body of research on the topic suggests that such schemes are not, in fact, as they seem.

References

Baron, J. (2008). *Thinking and deciding* (4th ed.). New York: Cambridge University Press.

Baron, J., Scott, S., Fincher, K., & Metz, S. E. (2015). Why does the cognitive reflection test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition, 4*(3), 265–284. https://doi.org/10.1016/j.jarmac.2014.09.003.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182. https://doi.org/10.1037/0022-3514.51.6.1173.

Barnes, A. (2016). Making intelligence analysis more intelligent: Using numeric probabilities. *Intelligence and National Security, 31*(3), 327–44. https://doi.org/10.1080/02684527.2014.994955.

Berry, D. C., Michas, I. C., Gillie, T., & Forster, M. (1997). What do patients want to know about their medicines, and what do doctors want to tell them?: A comparative study. *Psychology & Health, 12*(4), 467–480. https://doi.org/10.1080/08870449708406723.

Beyth-Marom, R. (1982). How probable is probable? A numerical translation of verbal probability expressions. *Journal of Forecasting, 1*(3), 257–269. https://doi.org/10.1002/for.3980010305.

Bilker, W. B., Wierzbicki, M. R., Brensinger, C. M., Gur, R. E., & Gur, R. C. (2014). Development of abbreviated eight-item form of the Penn Verbal Reasoning Test. *Assessment, 21*(6), 669–678. https://doi.org/10.1177/1073191114524270.

Bruine de Bruin, W., Fischbeck, P. S., Stiber, N. A., & Fischhoff, B. (2002). What number is "fifty-fifty"?: Redistributing excessive 50% responses in elicited probabilities. *Risk Analysis, 22*(4), 713–723. https://doi.org/10.1111/0272-4332.00063.

Bruine de Bruin, W., Fischhoff, B., Millstein, S. G., & Halpern-Felsher, B. L. (2000). Verbal and numerical expressions of probability: "It's a fifty-fifty chance." *Organizational Behavior and Human Decision Processes, 81*(1), 115–131. https://doi.org/10.1006/obhd.1999.2868.

Brun, W., & Teigen, K. H. (1988). Verbal probabilities: Ambiguous, context-dependent, or both? *Organizational Behavior and Human Decision Processes*, *41*(3), 390–404. https://doi.org/10.1016/0749-5978(88)90036-2.

Budescu, D. V., Broomell, S., & Por, H. (2009). Improving communication of uncertainty in the reports of the intergovernmental panel on climate change. *Psychological Science, 20*(3), 299–308. https://doi.org/10.1111/j.1467-9280.2009.02284.x.

Budescu, D. V., Por, H., & Broomell, S. B. (2012). Effective communication of uncertainty in the IPCC reports. *Climatic Change, 113*(2), 181–200. https://doi.org/10.1007/s10584-011-0330-3.

Budescu, D. V., Por, H., Broomell, S. B., & Smithson, M. (2014). The interpretation of IPCC probabilistic statements around the world. *Nature Climate Change, 4*(6), 508–512.

https://doi.org/10.1038/nclimate2194.

Budescu, D. V., & Wallsten, T. S. (1985). Consistency in interpretation of probabilistic phrases. *Organizational Behavior and Human Decision Processes, 36*(3), 391–405. https://doi.org/10.1016/0749-5978(85)90007-X.

Chang, W., Berdini, E., Mandel, D. R., & Tetlock, P. E. (2018). Restructuring structured analytic techniques in intelligence. *Intelligence and National Security, 33*(3), 337–356. https://dx.doi.org/10.1080/02684527.2017.1400230.

Chang, W., Chen, E., Mellers, B., & Tetlock, P. (2016). Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments. *Judgment and Decision Making, 11*(5), 509–526.

Collins, R. N., & Mandel, D. R. (2019). Cultivating credibility with probability words and numbers. *Judgment and Decision Making, 14*(6), 683–695.

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin numeracy test. *Judgment and Decision Making, 7*(1), 25–47.

Debs, A., & Monteiro, N. P. (2014). Known unknowns: Power shifts, uncertainty, and war. *International Organization, 68*(1), 1–31. https://doi.org/10.1017/S0020818313000192.

Dhami, M. K., Mandel, D. R. (2020). Words or numbers? Communicating probability in intelligence analysis. *American Psychologist*. Advance online publication. https://doi.org/10.1037/amp0000637.

Dhami, M. K., & Mandel, D. R. (2021, January 1). Negative context reduces the discriminability of verbal probabilities. Retrieved from https://psyarxiv.com/ma7rb.

Dhami, M. K., & Wallsten, T. S. (2005). Interpersonal comparison of subjective probabilities: Toward translating linguistic probabilities *Memory & Cognition, 33*(6)*,* 1057–1068. https://dx.doi.org/10.3758/BF03193213.

Dieckmann, N. F., Gregory, R., Peters, E., & Hartman, R. (2017). Seeing what you want to see: How imprecise uncertainty ranges enhance motivated reasoning. *Risk Analysis*, *37*(3)**,** 471–486. https://doi.org/10.1111/risa.12639.

Dieckmann, N. F., Peters, E., & Gregory, R. (2015). At home on the range? Lay interpretations of numerical uncertainty ranges. *Risk Analysis*, *35*(7)**,** 1281–1295. https://doi.org/10.1111/risa.12358.

Erev, I., & Cohen, B. L. (1990). Verbal versus numerical probabilities: Efficiency, biases, and the preference paradox. *Organizational Behavior and Human Decision Processes*, *45*(1), 1–18. https://doi.org/10.1016/0749-5978(90)90002-Q.

European Food Safety Authority, Hart, A., Maxim, L., Siegrist, M., von Goetz, N., da Cruz, C., et al. (2018). Guidance on communication of uncertainty in scientific assessments. *EFSA Journal, 17*(1), 1–73. https://doi.org/10.2903/j.efsa.2019.5520.

Fagley, N. S. (1993). A note concerning reflection effects versus framing effects. *Psychological Bulletin, 113*(3), 451–452. https://doi.org/10.1037/0033-2909.113.3.451.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior*

Research Methods, *39*, 175–191. https://doi.org/10.3758/BF03193146.

Fischhoff, B., & Bruine de Bruin, W. (1999). Fifty-fifty = 50? *Journal of Behavioral Decision Making, 12*(2), 149–167. https://doi.org/10.1002/(SICI)1099-0771(199906)12:2<149::AID-BDM314>3.0.CO;2-J.

Friedman, J. A. (2019). *War and chance: Assessing uncertainty in international politics*. New York: Oxford University Press.

Funtowicz, S. O., & Ravetz, J. R. (1990). *Uncertainty and quality in science for policy.* Dordrecht, The Netherlands: Kluwer Academic Publishers.

Geurts, B., & Nouwen, R. (2007). 'At least' et al.: The semantics of scalar modifiers. *Language*, *83*(3), 533–559. https://doi.org/10.1353/lan.2007.0115.

Haran, U., Ritov, I., & Mellers, B. A. (2013). The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Judgment and Decision Making*, *8*(3), 188–201.

Harris, A. J. L., Corner, A., Xu, J., & Du, X. (2013). Lost in translation? Interpretations of the probability phrases used by the intergovernmental panel on climate change in China and the UK. *Climatic Change, 121*, 415–425. https://doi.org/10.1007/s10584-013-0975-1.

Ho, E. H., Budescu, D. V., Dhami, M. K., & Mandel, D. R. (2015). Improving communication of uncertainty in climate science and intelligence analysis. *Behavioral Science & Policy, 1*(2), 43–55. https://doi.org/10.1353/bsp.2015.0015.

Hollingworth, H. L. (1910). The central tendency of judgment. *Journal of Philosophy, Psychology and Scientific Methods, 7*(17), 461–469. https://doi.org/10.2307/2012819.

Kent, S. (1951). *Strategic intelligence for American world policy*. Princeton, NJ: Princeton University Press.

Kent, S. (1964). Words of estimative probability. *Studies in Intelligence, 8*(4), 49–65.

Kolesnik, K., Silska-Gembka, S., & Gierusz, J. (2019). 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*(1), 25–49. https://dx.doi.org/10.24818/jamis.2019.01002.

Knapp, P., Gardner, P. H., & Woolf, E. (2016). Combined verbal and numerical expressions increase perceived risk of medicine side-effects: A randomized controlled trial of EMA recommendations. *Health Expectations, 19*(2), 264–274. https://doi.org/10.1111/hex.12344.

Lenhardt, E. D., Cross, R. N., Krocak, M. J., Ripberger, J. T., Ernst, S. R., Silva, C. L. , & Jenkins-Smith, H. C. (2020). How likely is that chance of thunderstorms? A study of how National Weather Service forecast offices use words of estimative probability and what they mean to the public. *Journal of Operational Meteorology*, *8*(5), 64–78, https://doi.org/10.15191/nwajom.2020.0805..

Lewis, S. C., King, A. D., Perkins-Kirkpatrick, S. E., & Wehner, M. F. (2019). Toward calibrated language for effectively communicating the results of extreme event attribution

studies. *Earth's Future*, *7*(9), 1020–1026. https://doi.org/10.1029/2019EF001273.

Lichtenstein, S., & Newman, J. R. (1967). Empirical scaling of common verbal phrases associated with numerical probabilities. *Psychonomic Science, 9*(10), 563–564. https://doi.org/10.3758/BF03327890.

Ligertwood, A., & Edmond, G. (2012). Expressing evaluative forensic science opinions in a court of law. Law, *Probability and Risk, 11*(4), 289–302. https://doi.org/10.1093/lpr/mgs016.

Lipkus, I. M., & Peters, E. (2009). Understanding the role of numeracy in health: Proposed theoretical framework and practical insights. *Health Education & Behavior, 36*(6), 1065–1081. https://doi.org/10.1177/1090198109341533.

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making*, *21*(1), 37–44. https://doi.org/10.1177/0272989X0102100105.

Liu, D., Juanchich, M., & Sirota, M. (2020). Focus to an attribute with verbal or numerical quantifiers affects the attribute framing effect. *Acta Psychologica, 208,* 103088. https://doi.org/10.1016/j.actpsy.2020.103088.

Mandel, D. R. (2007). Toward a concept of risk for effective military decision making. *DRDC Toronto technical report 2007–124*. Toronto: Defence Research and Development Canada.

Mandel, D. R. (2014). Do framing effects reveal irrational choice? *Journal of Experimental Psychology: General, 143*(3), 1185–1198 https://doi.org/10.1037/a0034207.

Mandel, D. R. (2015a). Accuracy of intelligence forecasts from the intelligence consumer's perspective. *Policy Insights from the Behavioral and Brain Sciences, 2*(1), 111–120. https://doi.org/10.1177/2372732215602907.

Mandel, D. R. (2015b). Instruction in information structuring improves Bayesian judgment in intelligence analysts. *Frontiers in Psychology, 6*, article 387. https://doi.org/10.3389/fpsyg.2015.00387.

Mandel, D. R., Collins, R. N., Risko, E. F., & Fugelsang, J. A. (2020). Effect of confidence interval construction on judgment accuracy. *Judgment and Decision Making*, *15*(5), 783–797.

Mandel, D. R., Dhami, M. K., Tran, S., & Irwin, D. (2021). Arithmetic computation with probability words and numbers. *Journal of Behavioral Decision Making*. Advance online publication. https://doi.org/10.1002/bdm.2232.

Mandel, D. R., & Irwin, D. (2020). Uncertainty, intelligence, and national security decisionmaking. *International Journal of Intelligence and CounterIntelligence*. Advance online publication. https://doi.org/10.1080/08850607.2020.1809056.

Mandel, D. R., & Irwin, D. (2021). 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*(3), e0248424. https://doi.org/10.1371/journal.pone.0248424.

Mandel, D. R., Wallsten, T. S., & Budescu, D. V. (2021). Numerically bounded linguistic probability schemes are unlikely to communicate uncertainty effectively. *Earth's Future, 9*, e2020EF001526. https://doi.org/10.1029/2020EF001526.

Marchio, J. (2014). "If the weatherman can...": The intelligence community's struggle to express analytic uncertainty in the 1970s. *Studies in Intelligence, 58*(4), 31–42.

Marvin, F. F. (2020). From BOGSAT to TurboTeam: Collaboration for national security teams in the age of analytics. In N. M. Scala and J. Howard (Eds.), *Handbook of military and defence operation research,* (pp. 283–304). Boca Raton, FL: CRC Press.

Mastrandrea, M. D., Mach, K. J., Plattner, G., & Matschoss, P. R. (2011). The IPCC AR5 guidance note on consistent treatment of uncertainties: A common approach across the working groups. *Climatic Change, 108*(4), 675–691. https://doi.org/10.1007/s10584-011-0178-6.

McQuiston-Surrett, D., & Saks, M. J. (2008). Communicating opinion evidence in the forensic identification sciences: Accuracy and impact. *Hastings Law Journal, 59*(5), 1159.

Mellers, B. A., Baker, J. D., Chen, E., Mandel, D. R., & Tetlock, P. E. (2017). How generalizable is good judgment? A multi-task, multi-benchmark study. *Judgment and Decision Making*, *12*(4), 369–381.

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., Bishop, M. M., Horowitz, M., Merkle, E., & Tetlock, P. (2015). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology: Applied, 21*(1), 1–14. https://doi.org/10.1037/xap0000040.

Moore, R. E., Kearfott, R. B., & Cloud, M. J. (2009). *Interval analysis*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Morgan, M. G. (1998). Commentary: Uncertainty analysis in risk assessment. *Human and Ecological Risk Assessment*, *4*(1), 25–39. https://doi.org/10.1080/10807039.1998.11009680.

Mullet, E., & Rivet, I. (1991). Comprehension of verbal probability expressions in children and adolescents. *Language and Communication*, *11*(3), 217–225. https://doi.org/10.1016/0271-5309(91)90007-I.

North Atlantic Treaty Organization. (2016). *AJP-2.1, edition B, version 1: Allied joint doctrine for intelligence procedures*. Brussels, Belgium: NATO Standardization Office.

Office of the Director of National Intelligence. (2015). *Intelligence community directive 203: Analytic standards*. Washington, DC: author.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45*(4), 867–872. https://doi.org/10.1016/j.jesp.2009.03.009.

Patt, A. G., & Dessai, S. (2005) Communicating uncertainty: lessons learned and suggestions for climate change assessment. *Comptes Rendus Geoscience, 337*(4), 425–441. https://doi.org/10.1016/j.crte.2004.10.004.

Patt, A. G., & Schrag, D. P. (2003). Using specific language to describe risk and probability. *Climatic Change, 61,* 17–30. https://doi.org/10.1023/A:1026314523443.

Poulton, E. C. (1994). *Behavioral decision theory: A new approach.* New York: Cambridge University Press.

Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin, 135*(6), 943–973. https://doi.org/10.1037/a0017327.

Stheeman, S. E., Mileman, P. A., van't Hof, M. A., & van der Stelt, P. F. (1993). Blind chance? An investigation into the perceived probabilities of phrases used in oral radiology for expressing chance. *Dentomaxillofacial Radiology, 22*, 135–139. https://doi.org/10.1259/dmfr.22.3.8299832.

Teigen, K. H., & Brun, W. (1995). Yes, but it is uncertain: Direction and communicative intention of verbal probabilistic terms. *Acta Psychologica, 88*(3), 233–258. https://doi.org/10.1016/0001-6918(93)E0071-9.

Teigen, K. H., & Brun, W. (1999). The directionality of verbal probability expressions: Effects on decisions, predictions, and probabilistic reasoning. *Organizational Behavior and Human Decision Processes, 80*(2), 155–190. https://doi.org/10.1006/obhd.1999.2857.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2017). Real-world correlates of performance on heuristics and biases tasks in a community sample: Heuristics and biases tasks and outcomes. *Journal of Behavioral Decision Making, 30*(2), 541–554. https://doi.org/10.1002/bdm.1973.

Stevens, S. S., & Greenbaum, H. B. (1966). Regression effect in psychophysical judgment. *Perception & Psychophysics, 1*, 439–446. https://doi.org/10.3758/BF03207424.

Wallsten, T. S., Budescu, D. V., Zwick, R., & Kemp, S. M. (1993). Preferences and reasons for communicating probabilistic information in verbal or numerical terms. *Bulletin of the Psychonomic Society*, 31(2) 135–138. https://doi.org/10.3758/BF03334162.

Wallsten, T. S., Fillenbaum, S., & Cox, J. A. (1986). Base rate effects on the interpretations of probability and frequency expressions. *Journal of Memory and Language*, *25*(5), 571–587. https://doi.org/10.1016/0749-596X(86)90012-4.

Weber, E. U., & Hilton, D. J. (1990). Contextual effects in the interpretations of probability words: Perceived base rate and severity of events. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(4), 781–789. https://doi.org/10.1037/0096-1523.16.4.781.

West, R. F., Toplak, M. E., & Stanovich, K. E. (2008). Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions. *Journal of Educational Psychology*, *100*(4), 930–941. https://doi.org/10.1037/a0012842.

Wintle, B. C., Fraser, H., Wills, B. C., Nicholson, A. E., & Fidler, F. (2019). Verbal probabilities: Very likely to be somewhat more confusing than numbers. *PLOS ONE, 14*(4), e0213522. https://doi.org/10.1371/journal.pone.0213522.

Yaniv, I., & Foster, D. P. (1995). Graininess of judgment under uncertainty: An accuracy-informativeness trade-off. *Journal of Experimental Psychology: General, 124*(4), 424–432. https://doi.org/10.1037/0096-3445.124.4.424.

Ziegler, D. K., Mosier, M. C., Buenaver, M., & Okuyemi, K. (2001). How much information about adverse effects of medication do patients want from physicians? *Archives of Internal Medicine*, *161*(5), 706–713. https://doi.org/10.1001/archinte.161.5.706.