Supplemental Material

Improving the communication of uncertainty in climate science and intelligence analysis

Emily H. Ho, David V. Budescu, Mandeep K. Dhami, & David R. Mandel

In this Supplemental Material, we provide a more detailed report of our two empirical studies, including our methods and more detailed results of our analyses and results. The first study is concerned with the communication of uncertainty in the context of describing climate change science. We derive an optimal phrase-to-numerical-range mapping of the verbal probabilities used in this community that results in maximal consensus and understanding between communicators. We show that our evidence-based lexicon is more effective than the lexicon used in the reports of the Intergovernmental Panel on Climate Change (IPCC). More specifically, our approach partitioned the [0, 1] probability range into four mutually exhaustive and exclusive intervals associated with four distinct phrases. The second study applies a similar approach to communicating uncertainty in intelligence analysis among analysts of various levels of experience. We show that favoring evidence-based approaches over existing organizational guidelines also improves communication among intelligence analysts.

Study 1: An Uncertainty Lexicon for Climate Change

The IPCC's translation aid (see columns 1 and 2 in Table 1) for communicating uncertainty has met with some criticism over its efficacy.[1,2] In Study 1, we developed an evidence-based verbal probability lexicon using people's empirical interpretations of probability phrases in the context of climate change. We assessed the effectiveness of these lexicons by comparing their performance to the IPCC's prescribed lexicon, using two independent samples.

Table 1. IPCC lexicon for translation of probability phrases and two evidence-based lexicons (Study 1)

Phrase	IPCC likelihood	Evidence-based method	
		Peak value	Membership function
Virtually certain	>99	—	—
Very likely	>90	65–100	75–100
Likely	>66	45–65	40–75
About as likely as not	33–66	—	—
Unlikely	<33	15–45	15–40
Very unlikely	<10	0–15	0–15
Exceptionally unlikely	<1	—	—

Note. All numbers are in percentages. Dashes indicate ranges. Dashes without numbers indicate phrases that the study did not examine due to how infrequently they appear in the IPCC.

Methods

Participants. The data were collected as part of a large international survey of over 11,000 respondents in 25 samples from 24 countries speaking 18 languages (see details in reference 3). The calibration sample used to derive the lexicon comprised U.S. participants, and the validation samples used to compare the evidence-based and the existing lexicons were taken from the United Kingdom and Australia. Table 2 characterizes the respondents in the three samples.

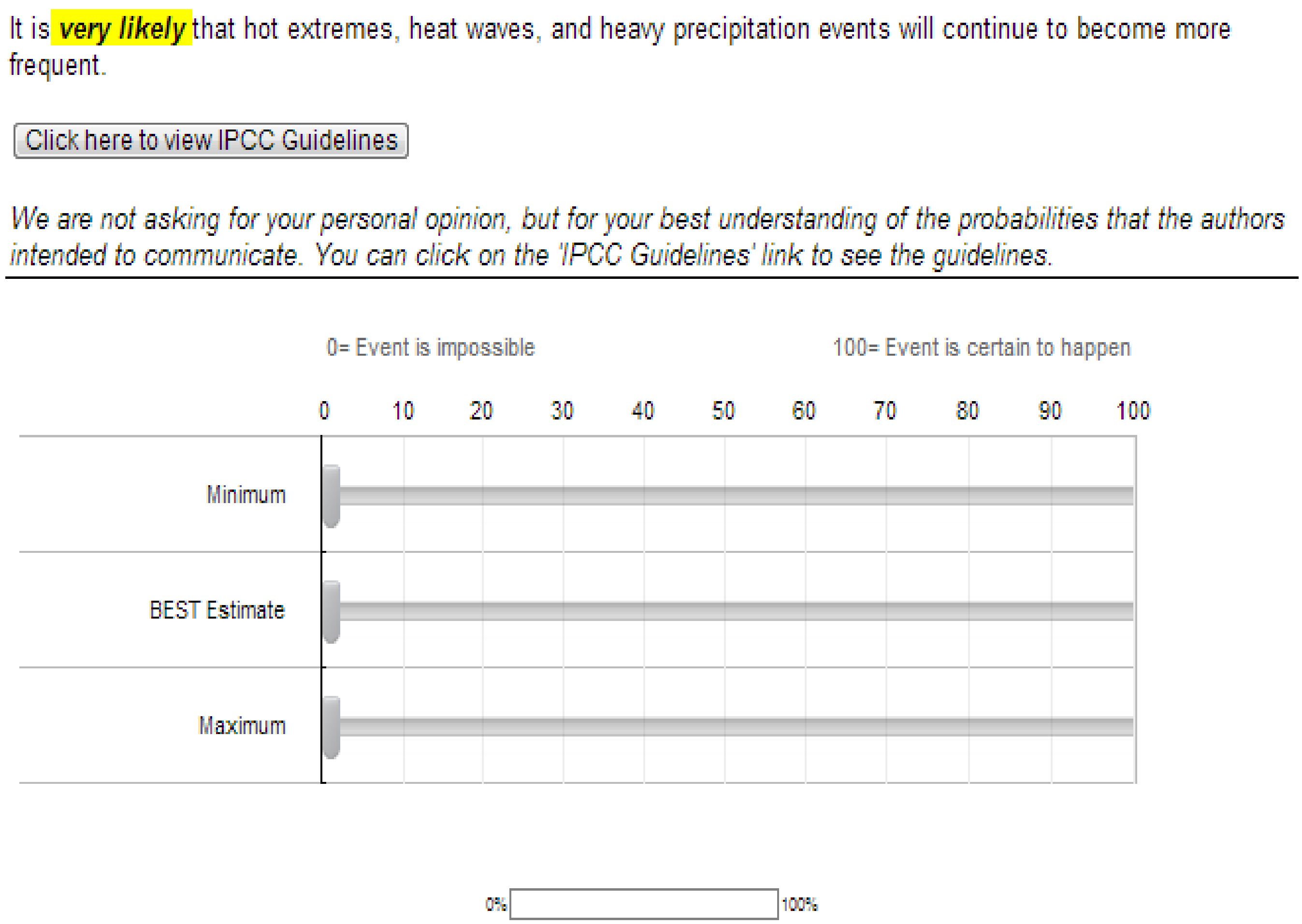
Table 2. Summary statistics for calibration and validation samples (Study 1)

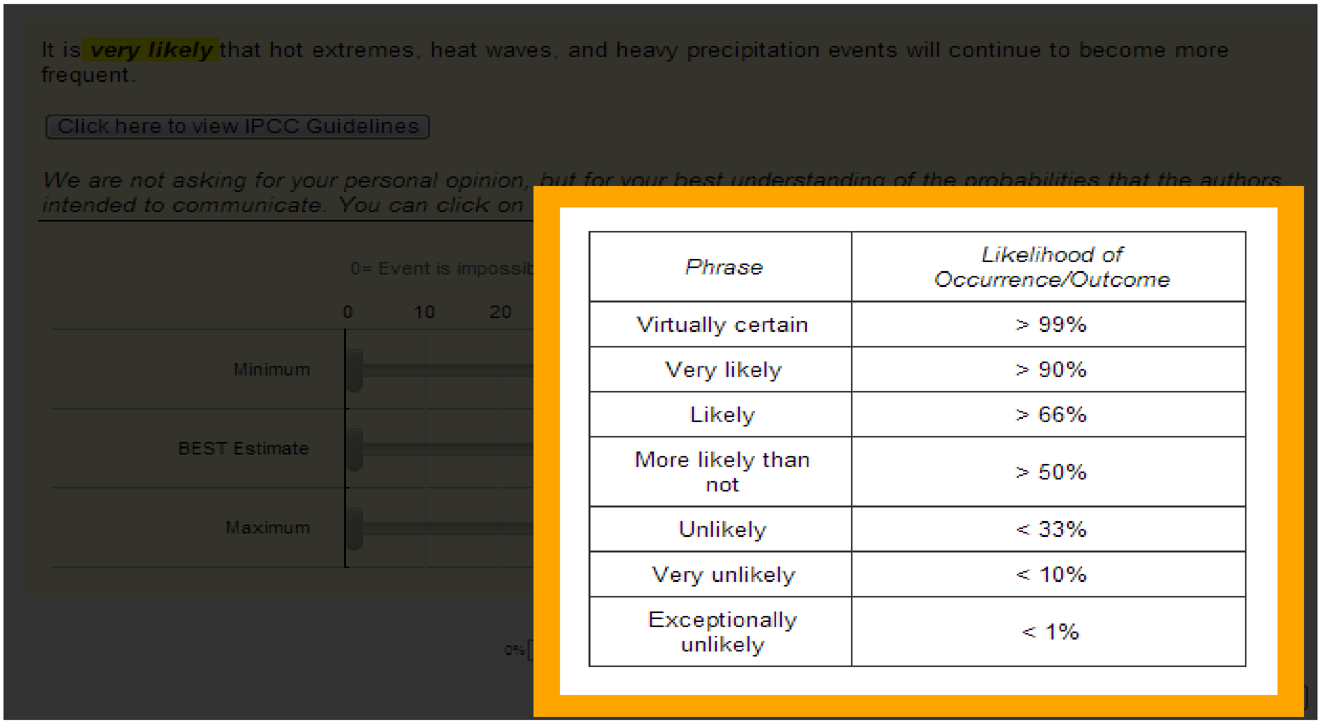
Sample	Country	Sample size for phrase				Summary statistics	
		Very unlikely	Unlikely	Likely	Very likely	Female	*M* age in years (*SD*)
Calibration	United States	331	345	352	347	52.4%	46.2 (14.3)
Validation	United Kingdom	162	169	164	177	50.0%	44.7 (12.2)
Validation	Australia	198	200	210	217	51.5%	49.9 (14.9)

General procedure. Participants were asked to complete two tasks.

1. *Evaluation of IPCC statements:* Participants read eight sentences from IPCC reports—two sentences for each of the four probability phrases (i.e., *very unlikely, unlikely, likely, very unlikely*) most frequently used in IPCC reports and thus the most relevant for policy efforts (see Table 1)[4]—and were asked to provide their lower-bound estimates (*L*), best estimates (*B*), and upper-bound estimates (*U*) of the phrase's intended meaning as it appears in the sentence (see Figure 1 for screenshot). Participants had access to the IPCC's conversion table when reading the sentences, visible in the second screenshot in Figure 1.

Figure 1. Screenshot of first task





Phrase	*Likelihood of Occurrence/Outcome*
Virtually certain	> 99%
Very likely	> 90%
Likely	> 66%
More likely than not	> 50%
Unlikely	< 33%
Very unlikely	< 10%
Exceptionally unlikely	< 1%

2. *Context-free estimation:* Participants saw the same four words without any context and were asked to indicate the values (*L*, *B*, and *U*) that they usually associate with these words.

We analyzed only those cases where the participant's best estimates were bracketed by their respective bounds (i.e., $L \leq B \leq U$).

Lexicon construction. We used the context-free estimates of the U.S. sample to construct two lexicons. We applied two evidence-based approaches—the *peak value* (PV) and the *membership function* (MF) methods—to associate each probability phrase with a range of numerical probabilities such that the resulting verbal lexicon maximized consensus across participants, resulting in two lexicons.

PV. The PV method relies on the empirical distributions of the best estimates (rounded to the nearest 0.05) of each phrase.[A] To evaluate the PV method, we used the best estimates of phrases and examined their distributions using the pair chart,[5] a nonparametric technique used to display and quantify the relationship between two distributions. We adapt the procedure to minimize the overlap between the distributions of the single best estimates for two phrases W_i and W_j (for example, *very unlikely* and *unlikely*), such that $W_i < W_j$.

Higher frequencies of probabilities imply a higher consensus across participants that a given numerical probability is an optimal descriptor of the probability phrase. The optimal cutoff points between adjacent phrases were obtained by evaluating the point of intersection between distributions of two adjacent phrases. The left panel of Figure 2 illustrates this process with the distributions of *very unlikely* and *unlikely*.

The MF method is based on the idea that probability phrases can be modeled as fuzzy subsets within the [0, 1] probability interval.[6,7] The MF describes how well a certain probability substitutes for the phrase (that is, how well a probability of .20 defines the phrase *very unlikely*). A membership of 0 denotes that the probability value does not define the phrase at all, whereas a membership of 1 indicates that a particular probability value absolutely exemplifies the phrase (that is, the peak of the MF). The MF method has proven to be a reliable and valid way to measure people's numerical understanding of probability phrases and their use of these phrases.[8–11]

In Study 1 we elicited only three values (*L*, *B*, and *U*), we interpolated triangular MFs linearly by setting Memb(*L*) = 0, Memb(*B*) = 1, and Memb(*U*) = 0:

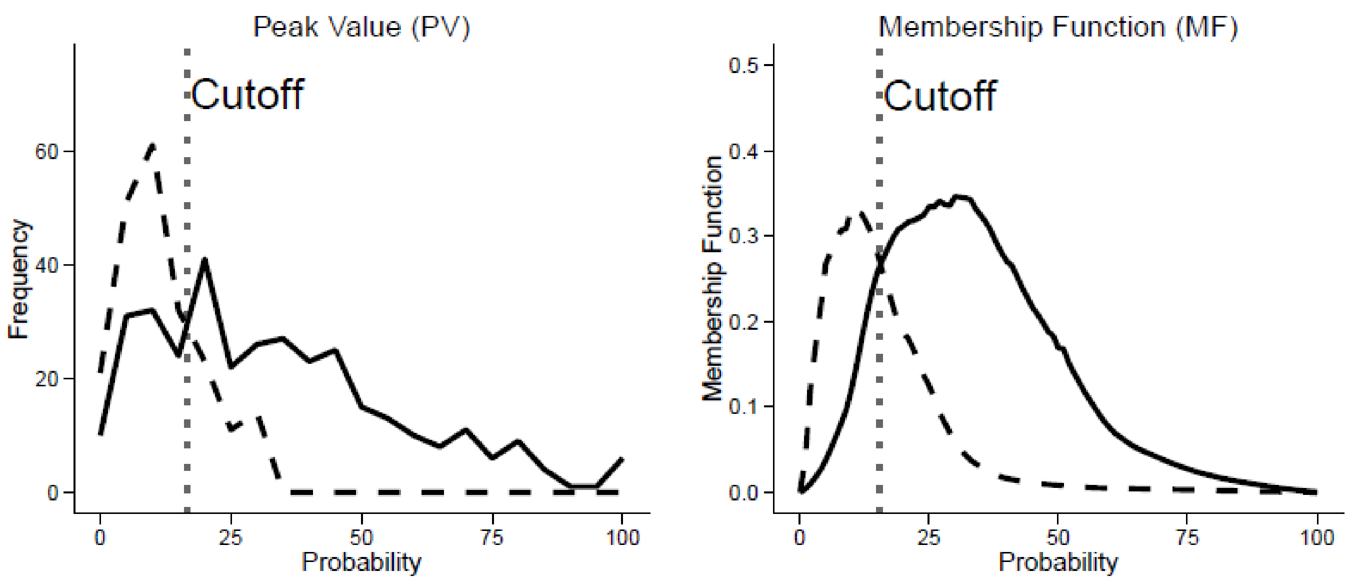
$$\square(\square) = \begin{cases} 0, \square \leq \square \\ \square(\square) = \dfrac{\square - \square}{\square - \square}, \square < \square \leq \square \\ \square(\square) = \dfrac{\square - \square}{\square - \square}, \square < \square \leq \square \\ 0, \square \geq \square \end{cases}$$

[A] Detection and trimming of extreme judgments for each phrase increases the statistical distance between adjacent phrases but negligibly affects the optimal thresholds. For a more technical consideration of various outlier detection methods used, see Appendix B.

In Study 2, MFs were linearly interpolated between adjacent points. Individual MFs were averaged over the probability interval [0,1] to obtain the group's MF for each phrase.[B]

The optimal cutoff points between adjacent phrases were obtained by identifying the range of values for which the (group) membership of a given phrase was higher than all other phrases. The right panel of Figure 2 illustrates this process with the group MFs of *likely* and *very likely*.

Figure 2. Illustration of determination of optimal cutoff points between two adjacent phrases using the peak value and membership function methods (Study 1)

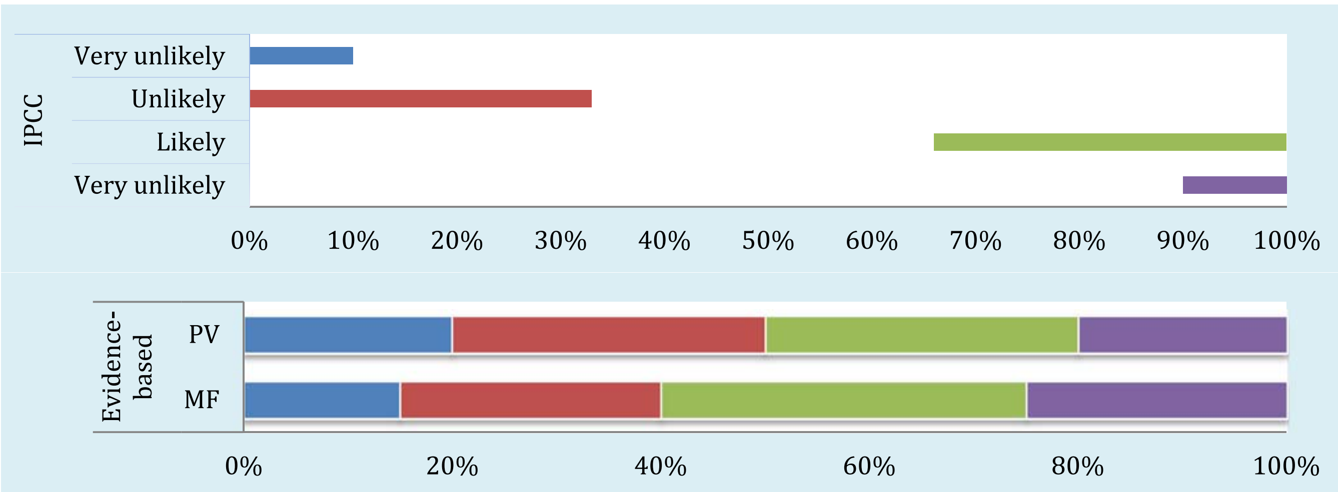


Data, Analyses, and Results

Table 1 displays the cutoffs prescribed by the IPCC lexicon and those inferred from our analyses and Figure 3 shows a comparison of these cutoffs. The most salient finding is that the IPCC's ranges for *very unlikely* and *very likely* are much narrower and shifted toward the extremes of the probability scale compared to the readers' intuitive and natural understanding of these phrases.

[B] For simplification, we focused on and averaged only individual MFs of the 21 probabilities divisible by 0.05.

Figure 3. Comparison of bounds in the IPCC guidelines with bounds from the evidence-based lexicons (Study 1)



Validation. We compared the communicative effectiveness of our two evidence-based lexicons with the existing IPCC by evaluating the consistency of the respondents assessments of the terms in the context of IPCC statements, by using two samples from English speaking countries (Australia and the United Kingdom). *Consistency* refers to the extent to which a participant's best estimate of a phrase complies (that is, falls within the range of the phrase) with either the IPCC or the two evidence-based cutoff points. A consistency rate of 1 indicates that all of the best estimates of a phrase fell within the specified ranges for that phrase; a consistency rate of 0 indicates that none of the best estimates of the phrase fell within the specified range; intermediate values identify the extent of consistency (for example, a consistency rate of .3 indicates that 30% of the participants' estimates fell within the specified range).

For each validation sample, consistency rates were calculated for each phrase and then averaged over all phrases to produce an overall lexicon consistency rate. The two evidence-based lexicons that we propose yielded mean consistency rates of 40% for the PV method, with 95% confidence interval (CI) [36.85, 45.91], and 43% for the MF method, with 95% CI [38.67, 47.89]. These rates clearly outperformed the current IPCC lexicon in both validation samples, which for the U.K. sample was 27%, with 95% CI [23.87, 33.13], and for the Australian sample was 25%, with 95% CI [28.65, 37.55].

Summary. Previous attempts to minimize the vagueness inherent in uncertainty communications in the domain of climate change science and policy have not been informed by rigorous empirical research. We derived optimal cutoffs for a verbal probability lexicon using the terms *very unlikely, unlikely, likely,* and *very likely* that were grounded in people's understanding of these phrases. When asked to interpret these probability phrases embedded in a climate change context, the two validation samples showed, as predicted, much higher consistency with our lexicons (see columns 3 and 4 of Table 1) than with the current IPCC guidelines, even though participants had access to the translation table used by the IPCC when interpreting them.

Study 2: A Lexicon for Communicating Uncertainty in Intelligence Assessments

We retested the efficacy of the methods proposed in Study 1 in the intelligence analysis domain with relevant domain experts. In this study, we used a larger number of phrases in developing a lexicon and a different method for eliciting the MFs of those phrases, permitting a test of the robustness of the evidence-based approach. Mean MFs were calculated for all the phrases. The PV optimal cutoffs—numerical bounds between two adjacent verbal phrases—were calculated in the same manner as in Study 1. For the PV method, the peak of the MF (the probability at which the group's mean MF had a maximal value) was the estimate. The MF optimal cutoffs were the intersection points of the mean MFs of adjacent terms.

An important aim of intelligence analysis is to provide probabilistic forecasts for future uncertain events (for example, 9/11, the Arab Spring). A calibration study[12] of 1,514 strategic intelligence forecast found that in forecasts in which uncertainties were numerically quantified, only 29.5% implied certainty (that is, probabilities of event occurrence equal to either 0 or 1). Miscommunicating or misunderstanding uncertainty in an analytic conclusion can lead to biased or erroneous decision making. For instance, in a study of managers of intelligence organizations, one manager stated, "There is a huge problem of language used to convey probability and importance/magnitude in terms of what the expressions mean to different people" (p. 23).[13,14]

After 9/11, there was increasing pressure for analysts writing intelligence products to be more explicit about uncertainties surrounding their judgments.[15] The National Intelligence Council (NIC) developed a rank-ordered verbal probability lexicon for communicating uncertainty (see Table 3).[16] The U.K. Defence Intelligence (DI) developed a different six-category lexicon in which phrases were translated into numerical ranges, although there are some gaps.[17,18] Neither lexicon was based on systematic empirical research, and empirical validation of their communicative effectiveness has yet to be ascertained.

Table 3. National Intelligence Council (NIC) and Defence Intelligence (DI) standardized probability lexicons (Study 2)

NIC		DI	
Phrase	Numerical value (%)	Phrase	Numerical value (%)
Remote	1–5	Remote/highly unlikely	<10
Very unlikely	5–20	Improbable/unlikely	15–20
Unlikely	20–45	Realistic possibility	25–50
Even chance	45–55		
Probably/likely	55–80	Probable/ likely	55–70
Very likely	80–95	Highly probable/ very likely	75–85
Almost certainly	95–99	Almost certain	>90

Methods & Analysis

We applied the same evidence-based methods developed in Study 1 and constructed a verbal probability lexicon for both the NIC and the DI. We also assessed the communicative effectiveness of the existing lexicons, examining consistency as well as the prescribed rank order and potential phrase equivalence (for example, *likely/probably*). A sample of Canadian military intelligence analysts was used to derive the optimal cutoffs of the evidence-based lexicons, and a U.K. sample of intelligence analysts was used to compare the consistency rate of the derived lexicon with that of the existing NIC and DI lexicons.

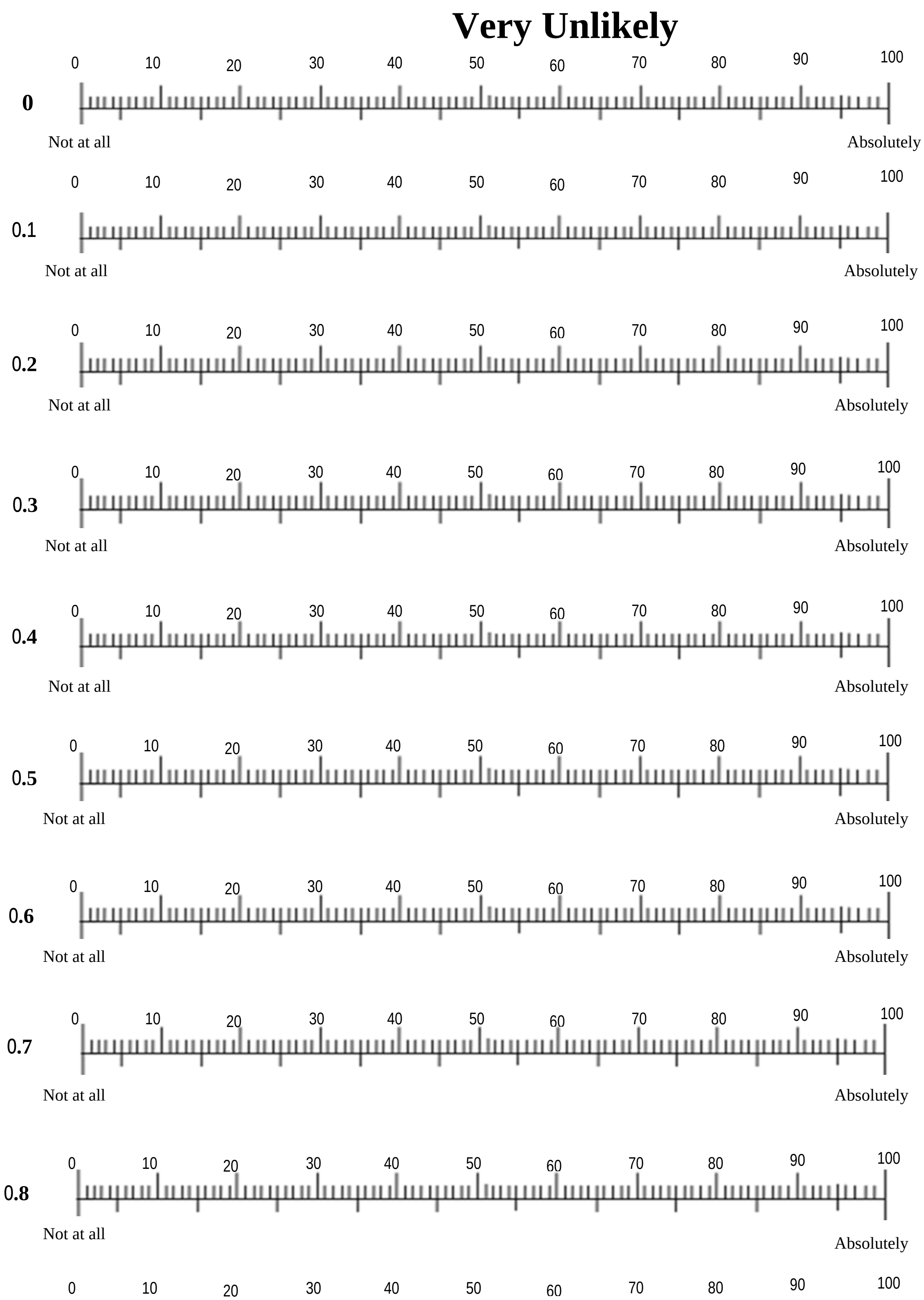
Methods

Participants. The calibration sample consisted of 34 intelligence analysts completing an intermediate intelligence officer course. The entry requirement for the course is a bachelor's degree. They were administered a survey to elicit the MFs of eight phrases taken from the NIC lexicon. All of these phrases appear in the DI lexicon as well. The validation sample consisted of 27 U.K. intelligence analysts who were administered a similar task (see reference 17, Study 1). Due to an administrative error, *very unlikely* judgments were not collected. We use the phrase *highly unlikely* as a substitute for *very unlikely* in our validation analysis.

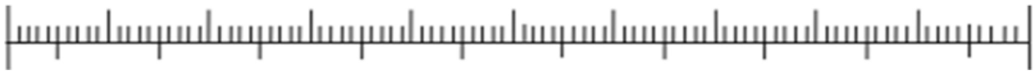
Procedure. Analysts rated the degree to which specific probabilities (expressed as percentages from 0% to 100% in increments of 10%) can substitute for probability phrases. Ratings were on a scale from 0, representing *not at all,* to 100, for *absolutely* (in 5-point increments), with each 10th point labeling an interval. Thus, each participant provided 11 ratings for each phrase. This is a refinement of the 3-point triangular MF interpolation method of Study 1. Figure 4 presents this elicitation method[19] for one phrase. The optimal cutoffs—the numerical bounds between two adjacent verbal phrases—were calculated in the same manner as in Study 1.

For the PV method, the peak of the MF (the probability at which an individual's MF has its maximal value) served as the estimate for the phrase. When individual responses yielded multiple peaks in the 11-point MF, the mean of these values was taken to obtain the PV. For the MF method, 11-point MFs were elicited from each participant for each phrase and averaged across all participants to obtain the group MFs for each phrase.

Figure 4. Sample elicitation display for Study 2

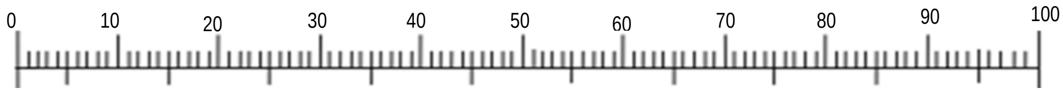


0.9



Not at all Absolutely

1.0



Not at all Absolutely

Data, Analyses, and Results

Rank order and phrase equivalence. We examined summary statistics for all of the phrases in the NIC lexicon. Table 4 presents the mean peak MFs for the phrases in both the calibration and the validation samples. It is reassuring that the order of both sets of mean estimates is consistent with the implied order of the phrases in the NIC and DI lexicons.

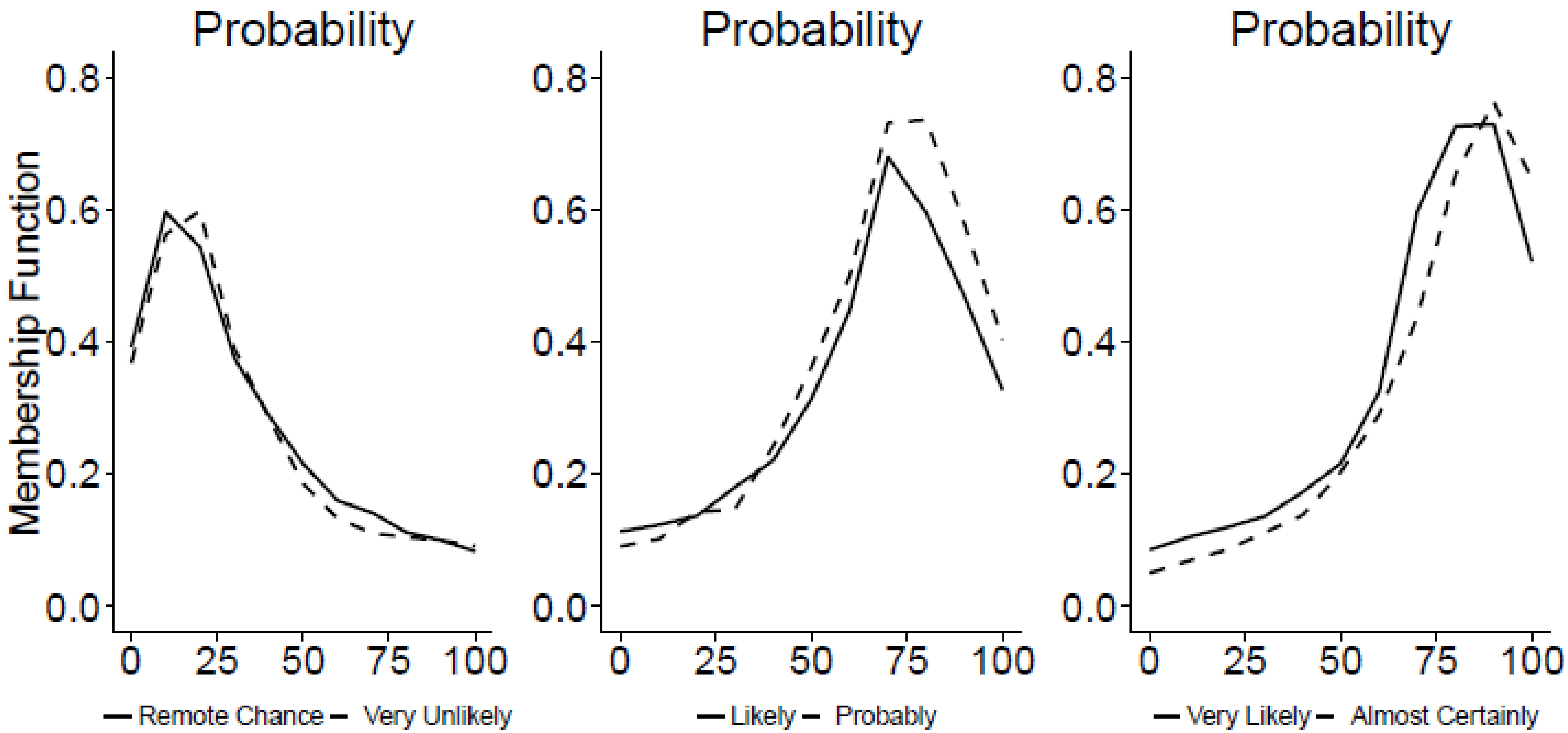
Table 4. Peaks of group MFs (and standard deviations) for all probability phrases (Study 2)

Sample		Remote chance	Very unlikely	Unlikely	Even chance	Likely	Probably	Very likely	Almost certainly
Calibration (Canada)	*M*	28.0	25.5	36.5	54.8	73.3	75.0	86.00	88.1
	SD	31.5	26.0	27.8	18.2	18.7	19.5	11.3	15.1
	N	32	32	31	30	32	22	31	32
Validation (United Kingdom)	*M*	16.8	18.2	28.3	50.0	74.4	78.8	82.9	88.6
	SD	18.1	21.2	22.4	—	9.8	9.7	8.4	16.7
	N	25	25	24	25	25	25	24	25

Note. In the validation sample, the phrase *highly unlikely* is a proxy for the phrase *very unlikely*.

The NIC and DI lexicons deem certain phrases to be interchangeable, meaning they can be used to represent the same numerical range. We calculated phrase differentiation between interchangeable phrases in the DI lexicon to assess this assumption. Figure 5 presents the average MFs of three phrase pairs in the NIC and DI lexicons—*remote chance/very unlikely*, *probably/likely,* and *very likely/almost certain*—and confirms that they are numerically indistinguishable and thus can be treated as synonyms. In light of this result, we use *probably* to represent *likely* in our evidence-based lexicons containing the full set of phrases found in the NIC lexicon and assign the same probability range to phrases that either the NIC or the DI deems equivalent in our evidence-based lexicon.

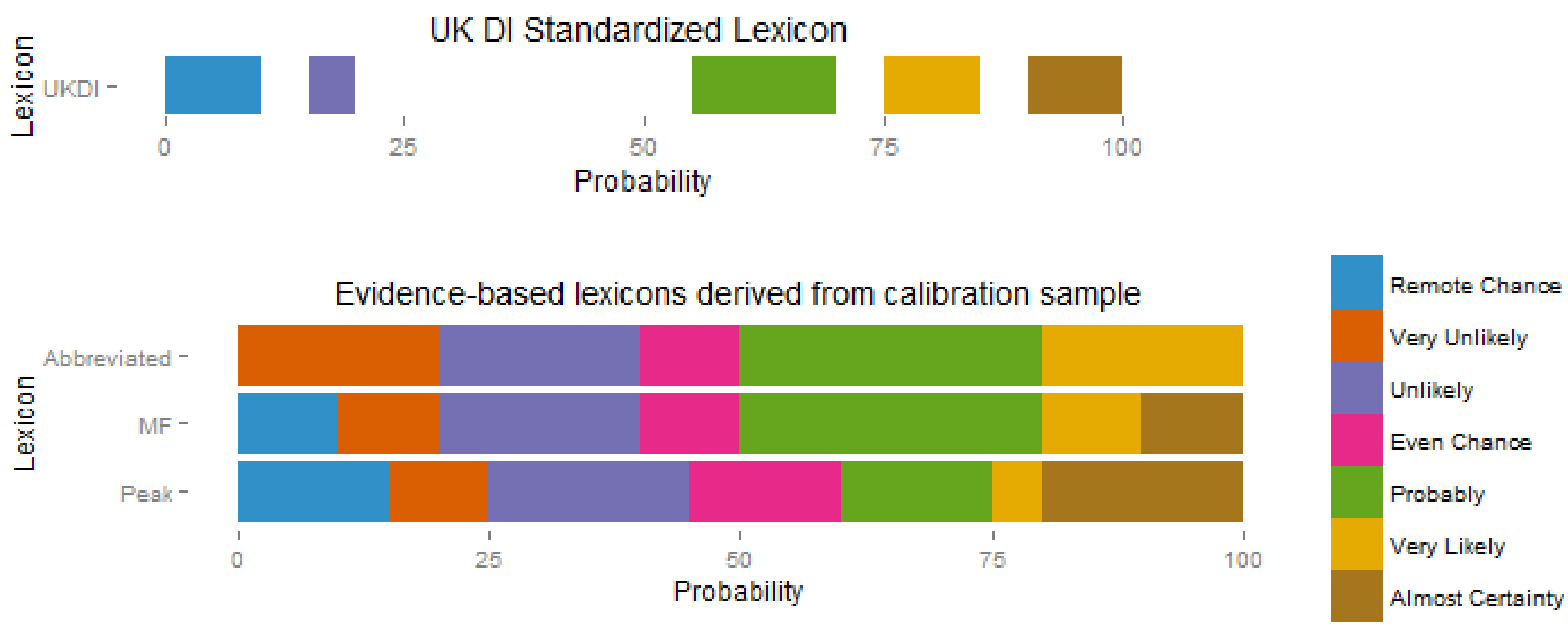
Figure 5. Comparison of average MFs for *remote chance* and *very unlikely* (left panel), *likely* and *probably* (middle panel), and *very likely* and *almost certainly* (right panel) in the Canadian sample (Study 2)



Probability
Probability
Probability
Membership Function
0.8
0.6
0.4
0.2
0.0
0
25
50
75
100
Remote Chance
Very Unlikely
Likely
Probably
Very Likely
Almost Certainly

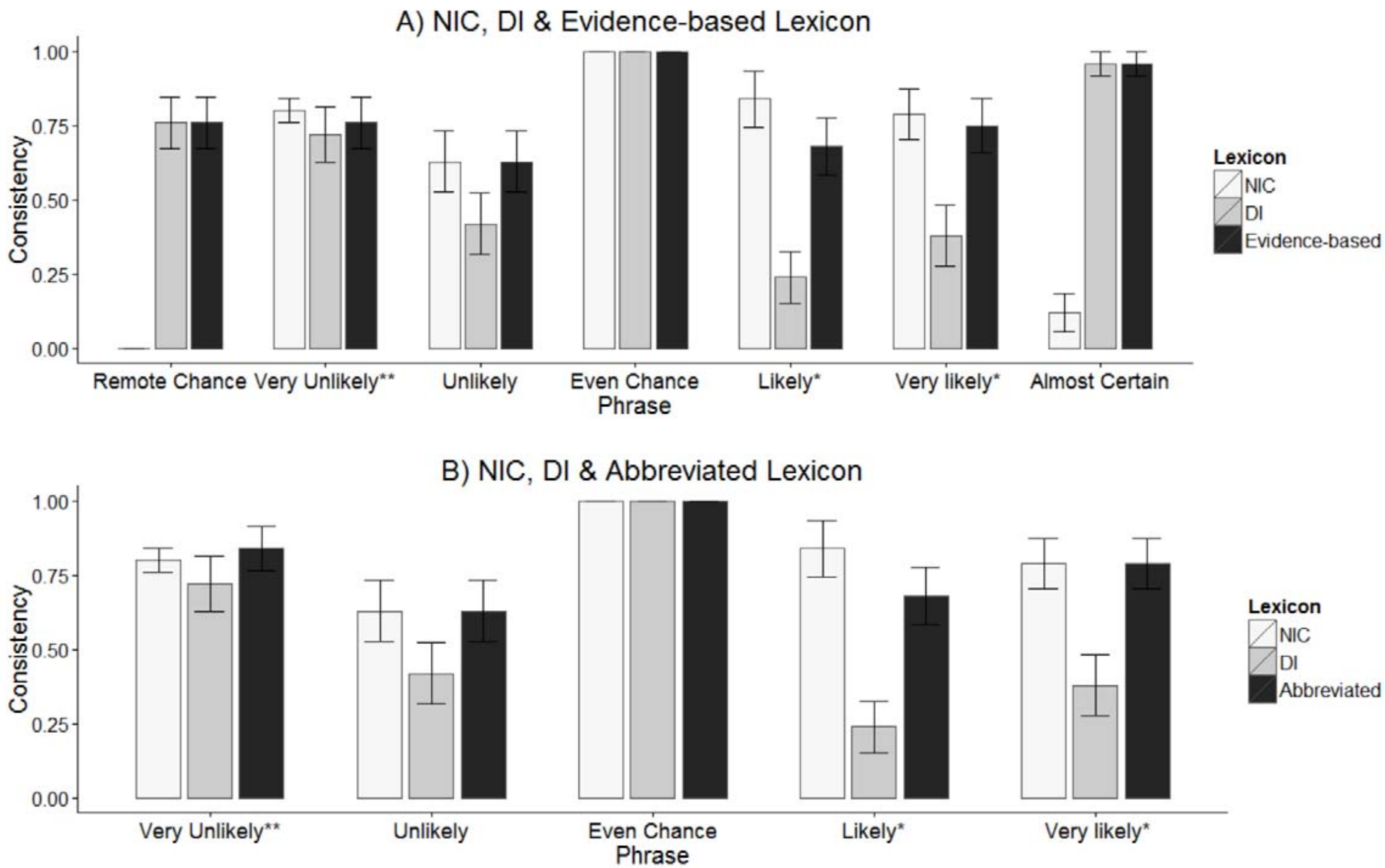
Evidence-based lexicon cutoffs. Figure 6 shows that the phrase ranges in the DI lexicon are generally narrower than the two evidence-based lexicons calibrated by the Canadian sample. Since the NIC lexicon does not include numerical equivalents, we cannot perform this analysis for the NIC lexicon.

Figure 6. Abbreviated and optimal thresholds at 10% trim compared with U.K. Defence Intelligence (DI) lexicon (Study 2)



Abbreviated lexicon. Can a simplified lexicon containing fewer but most frequently used terms provide a better differentiation between adjacent phrases and more effective communication? Our simplified lexicon omits *remote chance* and *almost certain*. Figure 5 compares the consistency rates (that is, percentage of cases in which the participants' judgments fell within either the evidence-based or the organizational lexicons) between the DI lexicon and (a) the complete evidence-based lexicon and (b) the abbreviated lexicon. The U.K. analysts' judgments of the phrases in the DI lexicon showed high consistency for extreme phrases with clear directionality (e.g., *highly unlikely* and *almost certainly*), but they had low consistency for other phrases.

Figure 5. (a) Consistency rates of the National Intelligence Council (NIC) lexicon, Defence Intelligence (DI) lexicon, and full evidence-based lexicon (Study 2). (b) Consistency rates of NIC lexicon, DI lexicon, and abbreviated evidence-based lexicon (Study 2).



** Due to an administrative error, *very unlikely* judgments were not collected. We use the phrase *highly unlikely* as a substitute for *very unlikely* in our validation analysis.

* The phrase *likely* is shown to be numerically equivalent to *probably*, so the validation applies to the latter phrase as well.

* The phrase *very likely* is shown to be numerically equivalent to *highly probable*, so the validation applies to the latter phrase as well.

Table 5 presents the empirically derived lexicons with optimal cutoff points and consistency rates for the two methods (that is, PV and MF), for the complete lexicon (eight phrases) and Table 6 presents the abbreviated version (five phrases). For most phrases and, on average, within each empirically constructed lexicon, the intelligence analysts' judgments showed much higher consistency with the two evidence-based lexicons we developed than with the DI's existing lexicon. This advantage held for both the complete and the abbreviated lexicons and was more pronounced using the MF method.

Table 5. Consistency of judgments with NIC and DI lexicons and two full evidence-based lexicons (Study 2)

	Evidence-based lexicon				Consistency			
Phrase	NIC	DI	PV	MF	NIC *M* (*SD*)	DI *M* (*SD*)	PV *M* (*SD*)	MF *M* (*SD*)
Remote chance	1–5	0–10	0–15	0–10	0 (0.00)	0.76 (0.44)	0.76 (0.44)	0.76 (0.44)
Very unlikely[a]	5–20	0–10	15–25	10–20	0.80 (0.20)	0.72 (0.46)	0.12 (0.33)	0.76 (0.44)
Unlikely	20–45	15–20	25–45	20–40	0.63 (0.51)	0.42 (0.50)	0.38 (0.49)	0.63 (0.49)
Realistic possibility	—	25-40	—	—	0.04 (0.20)	0	—	—
Even chance[b]	45–55	—	45–60	40–50	1.00 (—)	—	1.00 (—)	1.00 (—)
Probably	55–80	55–70	60–75	50–80	0.84 (0.37)	0.24 (0.44)	0.56 (0.51)	0.68 (0.48)
Very likely	80–95	75–85	75–80	80–90	0.79 (0.41)	0.38 (0.49)	0.29 (0.46)	0.75 (0.44)
Almost certain	95–99	90–100	80–100	90–100	0.12 (0.33)	0.96 (0.20)	0.96 (0.20)	0.96 (0.20)
M consistency					0.53	0.56[+]	0.58	0.79

Note. NIC = National Intelligence Council; DI = Defence Intelligence; PV = peak value; MF = membership function. Dashes represent ranges.

[a]In the validation sample, the phrase *highly unlikely* is a proxy for the phrase *very unlikely*.
[b]Although the term *even chance* is not part of this list, we included it in the mean calculation (assuming 100% consistency)Table 6. Consistency of judgments with two abbreviated evidence-based lexicons (Study 2)

	Evidence-based lexicon		Consistency			
	PV	MF	NIC *M* (*SD*)	DI *M* (*SD*)	PV*M* (*SD*)	MF*M* (*SD*)
Very unlikely[a]	0–15	0–10	0.80 (0.20)	0.72 (0.46)	0.80 (0.41)	0.84 (0.37)
Unlikely	15–45	10–40	0.63 (0.51)	0.42 (0.50)	0.79 (0.41)	0.63 (0.49)
Even chance	45–60	40–50	1.00 (—)	—	1.00 (—)	1.00 (—)
Probably	60–75	50–80	0.84 (0.37)	0.24 (0.44)	0.48 (0.51)	0.68 (0.48)
Very likely	75–100	80–100	0.79 (0.41)	0.38 (0.49)	0.83 (0.38)	0.79 (0.41)
M consistency			0.81	0.55[+]	0.78	0.79

[a]In the validation sample, the phrase *highly unlikely* is a proxy for the phrase *very unlikely*.

Note. The abbreviated lexicons exclude the phrases *remote chance* and *almost certain*. NIC = National Intelligence Council; DI = Defence Intelligence; PV = peak value, MF = membership function. Dashes represent ranges.

Summary. Study 2 replicated the key finding of Study 1 with a larger set of probability phrases and with estimates from two samples of experts in the intelligence domain. Specifically, both evidence-based lexicons (the abbreviated and the long versions) exhibited much higher consistency with participants' interpretations than with existing guidelines. This demonstrates that professionals operating in their domain of expertise can also benefit from systematic empirical research aimed at validating and improving existing organizational guidelines.

Supplemental References

1. Budescu, D. V., Por, H.-H., & Broomell S. B. (2012). Effective communication of uncertainty in the IPCC reports. *Climactic Change, 113,* 181–200.

2. Morgan, M. G., Adams, P. J., & Keith, D. W. (2006). Elicitation of expert judgments of aerosol forcing. *Climactic Change, 75,* 195–214.

3. Budescu, D. V., Por, H., Broomell, S., & Smithson, M. (2014). Interpretation of IPCC probabilistic statements around the world. *Nature Climate Change, 4,* 508–512. doi:10.1038/nclimate2194

4. Swart, R., Bernstein, L., Ha-Duong, M., & Petersen, A. (2009). Agreeing to disagree: Uncertainty management in assessing climate change, impacts and responses by the IPCC. *Climactic Change, 92,* 1–29.

5. Quade, D. (1973). The pair chart. *Statistica Neerlandica*. 27(1). 29-45. doi: 10.1111/j.1467-9574.1973.tb00206.x

6. Wallsten, T. S., Budescu, D. V., Rapoport, A., Zwick, R., & Forsyth, B. (1986). Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General, 115,* 348–365.

7. Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8,* 338–353.

8. Dhami, M. K. (2008). On measuring quantitative interpretations of reasonable doubt. *Journal of Experimental Psychology: Applied, 14,* 353–363.

9. Dhami, M. K., & Wallsten, T. S. (2005). Interpersonal comparison of subjective probabilities: Toward translating linguistic probabilities. *Memory & Cognition, 33,* 1057–1068.

10. Karelitz, T. M., & Budescu, D. V. (2004). You say “probable” and I say “likely”: Improving interpersonal communication with verbal probability phrases. *Journal of Experimental Psychology: Applied, 10,* 25–41.

11. Dhami, M. K., Lundrigan, S., & Mueller-Johnson, K. (2015). Instructions on reasonable doubt: Defining the standard of proof and the juror’s task. *Psychology, Public Policy, and Law, 21,* 169–178.

12. Mandel, D. R., & Barnes, A. (2014). Accuracy of forecasts in strategic intelligence. *Proceedings of the National Academy of Sciences, USA, 111,* 10984–10989.

13. Derbentseva, N., McLellan, L., & Mandel, D. R. (2010). *Issues in intelligence production: Summary of interviews with Canadian intelligence managers* (Technical Report ADA551144). Toronto, Ontario, Canada: Defence Research and Development Toronto.

14. Adams, B. A., Thomson, M., Derbentseva, N., & Mandel, D. R. (2011). Capability challenges in the human domain for intelligence analysis: Report on community-wide discussions with Canadian intelligence professionals. Tor Can Def Res Dev [Internet]. 2011; Available from: http://pubs.drdc-rddc.gc.ca/PDFS/unc118/p536570_A1b.pdf

15. Weiss, C. (2007). Communicating uncertainty in intelligence and other professions. *International Journal of Intelligence and CounterIntelligence,21,* 57–85.

16. National Intelligence Council. (2007). *Prospects for Iraq's stability: A challenging road ahead*. Retrieved from http://fas.org/irp/dni/iraq020207.pdf

17. Dhami, M. K. (2013). *Understanding and communicating uncertainty in intelligence analysis.* Unclassified report prepared for Her Majesty's Government, United Kingdom. (Available from the author)

18. Defence Intelligence. (n.d.). *Quick wins for busy analysts.* London, United Kingdom: Ministry of Defence.

19. Budescu, D. V., Karelitz, T. M., & Wallsten, T. S. (2003). Predicting the directionality of probability words from their membership functions. *Journal of Behavioral Decision Making,, 16,* 159–180.