Policy Insights from the
Behavioral and Brain Sciences
2015, Vol. 2(1) 111–120

DOI: 10.1177/2372732215602907
bbs.sagepub.com
SAGE

Accuracy of Intelligence Forecasts From the Intelligence Consumer's Perspective

David R. Mandel[1,2]

Abstract

Accurate forecasting is a vital part of intelligence assessment. Only recently has intelligence forecast accuracy been quantitatively tracked. Mandel and Barnes reported on a long-term study of intelligence forecasts that examined accuracy from the analysts' perspective using numerical probabilities that were not reported to intelligence consumers. The present research reassessed the accuracy of those forecasts from an intelligence consumer's perspective using findings from an experiment that elicited from subjects' numerical probability equivalents for the linguistic probabilities that consumers would have read in intelligence reports. Forecast accuracy was undiminished when assessed from the consumers' perspective (inferred from subjects' median numerical equivalents) because the intended meaning of the probability terms used by the intelligence unit corresponded well to the average meaning assigned by subjects. The findings also showed that interpretations of linguistic probabilities are context-dependent. Linguistic probabilities were discriminated better when applied to outcomes that represented successes rather than failures.

Keywords

forecasting, accuracy, linguistic probabilities, communication standards, intelligence analysis

Tweet

Long-term study of strategic intelligence shows good forecasting and evidence of effective communication of uncertainties to policymakers.

Key Points

- Forecasting is a vital function of intelligence organizations and forecasting accuracy can and should be quantitatively measured in a proactive manner.
- Most intelligence organizations use linguistic probabilities (e.g., *unlikely, probable, almost certain*) to communicate uncertainties in forecasts to intelligence consumers.
- Strategic intelligence forecasts can be highly accurate when assessed from the intelligence producers' viewpoint (using their numerical estimates) and from the intelligence consumers' viewpoint (using inferred probabilities).
- Linguistic probabilities are susceptible to context effects, which pose challenges to communicating uncertainties clearly.
- Behavioral research methods can play an important role in accounting for and promoting forecast accuracy and effectiveness in communicating uncertainties to decision-makers.

Introduction

> Better earlier warning than later mourning.
>
> Jewish proverb

> When men speak of the future, the gods laugh.
>
> Chinese proverb

A vital function of intelligence organizations is to communicate timely indicative information about the future to decision-makers. As the Jewish proverb above conveys, accurate indications about consequential future events that arrive early enough can help decision-makers avert trouble. In fact, the predictive function of intelligence is not only about forewarning and attending to a state's prevention goals. State leaders and policymakers also require predictive intelligence

[1] Defence Research and Development Canada, Toronto, Ontario, Canada
[2] York University, Toronto, Ontario, Canada

Corresponding Author:
David R. Mandel, Toronto Research Centre, Defence Research and Development Canada, 1133 Sheppard Avenue West, Toronto, Ontario, Canada M3K 2C9.
Email: david.mandel@drdc-rddc.gc.ca

to achieve promotion goals in multiple spheres of state policy, from effectively managing diplomatic relations to pursuing sound economic policies in a globalized economy. Yet, if the Jewish proverb supplies the raison d'être for predictive intelligence, the Chinese proverb that follows it lays bare a major challenge. The future, deeply uncertain in a myriad of respects, does not give itself up easily even when forecasters exert great effort.

Presumably, all parties with a vested interest in predictive intelligence—that is, both the producers and consumers of intelligence—would like to know how good analytic forecasts are. At least one study has shown that intelligence community (IC) personnel regard prediction as slightly more important than the descriptive function of intelligence analysis (Adams, Thomson, Derbentseva, & Mandel, 2012). Nevertheless, remarkably little is known about accuracy because most intelligence organizations do not systematically track forecast quality (Betts, 2007; Dhami, Mandel, Mellers, & Tetlock, in press; Friedman & Zeckhauser, 2012). In the United States, accountability processes promulgated by the Office of the Director of National Intelligence (ODNI) that touch on predictive accuracy are process-oriented guidelines that sound sensible but which are difficult to monitor. For instance, ODNI's (2015) IC Directive 203 on Analytic Standards states, "Analytic products should apply expertise and logic to make the most accurate judgments and assessments possible, based on the information available and known information gaps" (p. 4).

The emphasis on process accountability to the virtual exclusion of outcome accountability (e.g., tracking accuracy rates over time) is problematic because it is unclear how effective compliance with process guidelines is in making intelligence products better—that is, in the case of forecasting, improving skill components of accuracy such as calibration and discrimination. Nor do such guidelines allow intelligence organizations to maintain a scorecard of their forecasting performance, or to learn how such scores vary by putative moderators such as organization, level of experience, region of interest, level of assessment (e.g., tactical vs. strategic), to name but a few possibilities. Political leaders and the media often blame the IC for strategic surprises, promoting defensive maneuvers aimed at minimizing reoccurrences of the last big error (Tetlock & Mellers, 2011). In the absence of transparent, regular recordkeeping, it is difficult to contextualize a single intelligence failure, especially when it happens to be consequential—even if the failure happened to be near impossible to predict. Thus, it is in the IC's interest to verify its performance in more transparent ways.

Tracking the predictive accuracy of finished intelligence is doable, but the IC would have to invest the required resources. Processes would need to be developed and implemented that enable analysts or their managers to identify specific forecasts within reports, to record those forecasts in a database, to have the relevant outcomes tracked (i.e., for each forecasted event, did it occur or not?), to apply scoring rules appropriate to the data, to interpret the results, and to provide mechanisms whereby the overall activity could stimulate individual and organizational learning.

Alan Barnes, a former Canadian intelligence director, and I conducted a rare (possibly unique) example of this sort of monitoring exercise (Mandel, 2014, presents a summary). We studied the accuracy of strategic intelligence forecasts intended for an elite decision-making audience consisting of senior clients across government. We had subject-matter experts track the outcomes of the forecasted events, allowing us to examine approximately 6 years of forecasts from an intelligence unit reporting on the Middle East and Africa. Quantitative scoring relied on techniques commonly used to study forecasting skill, including mean Brier scores and their skill-component indices of calibration and discrimination (Yaniv, Yates, & Smith, 1991). Prior to these activities, however, analytic processes detailed in Barnes (2015) had to be implemented.

In brief, analysts in the unit were required to identify each forecast they made in their reports, distinguishing them from other types of assessments, such as causal inferences, or from purely descriptive statements. For each of the forecasts, analysts were further required to assign a numerical probability to the forecasted outcome using a 9-point scale corresponding to percentage chances of 0, 10, 20 to 30 (treated as 25 in the accuracy analyses), 40, 50, 60, 70 to 80 (likewise, treated as 75), 90, and 100. Those probabilities were used in scoring the skill components. Although those numerical probabilities represent real strategic intelligence forecasts (i.e., not merely research tasks using real analysts), they were recorded solely for internal review purposes. The finished intelligence reports sent to intelligence consumers included only linguistic probabilities that were also generated by the analysts and that accompanied their numerical probability forecasts. In other words, the analysts' words were disseminated, but not their numbers.

The findings reported in Mandel and Barnes (2014; see also Mandel, Barnes, & Richards, 2014) revealed very good calibration and discrimination skill in the 1,514 analytic forecasts examined for accuracy. For instance, the accuracy rate was 94%, and the mean Brier score (i.e., the mean squared deviation between the probabilistic forecast and the outcome, coded 0 for nonoccurrence and 1 for occurrence) was 0.074, with 76% of outcome variance explained by the forecasts. The calibration index, a measure of forecaster reliability, was 0.016 (0 represents perfect calibration), with most of the miscalibration due to underconfidence (cf. Tetlock, 2005), which was largely correctable using a mathematical transformation that made forecasts more extreme, on average.

Although Mandel and Barnes's (2014) study describes the forecasting skill of a strategic intelligence unit from the intelligence producers' perspective, that article does not address the issue of how accurate those forecasts would be when interpreted from a consumer's perspective. The question of how good the forecasts are from a consumer's perspective is an important one because intelligence is not an end in itself. As noted earlier, the purpose of intelligence is to reduce uncertainties, where warranted, in an effort to improve decision-making by political elites (Fingar, 2011; Scowcroft, 2009), and to at least accurately characterize uncertainties for those elites (Friedman & Zeckhauser, 2012).

Using the Mandel and Barnes (2014) forecast data set in conjunction with data from a new experiment on the interpretation of the linguistic probabilities used in those forecasts, the present research addressed the challenge of inferring forecast accuracy from the intelligence consumer's perspective. The importance of that challenge follows directly from the high skill level shown in the numerical probability forecasts. Had performance been dismal, one could easily have inferred that performance from a consumer's perspective would be equally dismal or worse because any loss of communication fidelity would impair rather than improve forecast accuracy. However, given that accuracy based on analysts' numerical probability forecasts was very good, accuracy could conceivably be worse from the consumer's perspective. That is, forecast accuracy from the intelligence consumer's perspective (relative to that of the producer) ought to deteriorate as miscommunication of the forecast probabilities increases.

Communicating Intelligence With Linguistic Probabilities

Sherman Kent, a former Central Intelligence Agency veteran who is often credited as the father of modern intelligence analysis, was among the first to write about the communication problems associated with using words to convey degrees of uncertainty or probability in intelligence assessments. After being asked by a policy official what was meant by "serious possibility" (of a Soviet attack on Yugoslavia) in a 1951 National Intelligence Estimate, Kent (1964) was jolted to find that his colleagues' interpretations of the term ranged from a 20% to an 80% chance. Kent attempted to tame the language of uncertainty by institutionalizing a standard whereby an organizationally approved set of terms was assigned numerical probability ranges. Although Kent's original standard was eventually abandoned in the U.S. IC, the general approach of trying to standardize the meaning of a lexicon of linguistic probabilities either by institutionalizing a rank ordering of terms (as the U.S. National Intelligence Council [NIC] does; Wheaton, 2012) or by associating numbers or numerical ranges to a set of terms (Barnes, 2015; Ho, Budescu, Dhami, & Mandel, in press) has been repeated elsewhere in ICs of various countries.

The use of words rather than numbers to convey uncertainty is unsurprising. People have a preference for communicating uncertainty to others using words, even if they prefer to receive such information in the form of numbers (Brun & Teigen, 1988; Murphy, Lichtenstein, Fischhoff, & Winkler, 1980; Wallsten, Budescu, Zwick, & Kemp, 1993). From the communicator's perspective, linguistic probabilities provide greater elasticity of meaning than numerical probabilities. Not only do words free creative expression and lend themselves to the narrative structure of intelligence reporting, they can also ease the short-term accountability pressures on analysts or other expert assessors. Indeed, some official documents, such as NIC reports, explain the IC's preference for linguistic probabilities in terms of wanting to avoid a false sense of precision (Wheaton, 2012), a view also expressed by many analysts (Barnes, 2015; Weiss, 2008), and one that reiterates the views of a Presidential-Congressional Commission on Risk Assessment and Risk Management (Morgan, 1998).

Yet, what strikes communicators as a desirable quality can also contribute to miscommunication. Like message senders, receivers assign probability ranges to linguistic probabilities, and those ranges exhibit considerable variability across individuals (e.g., Dhami & Wallsten, 2005; for a review, see Wallsten & Budescu, 1995), especially as samples become more heterogeneous in their demographic characteristics (Clarke, Ruffin, Hill, & Beamen, 1992). Nor does expertise provide an effective antidote to vagueness. In one study (Morgan, 1998), members of the Executive Committee of the Environmental Protection Agency's (EPA) Science Advisory Board were asked to provide quantitative range estimates of the terms *likely* and *not likely* (terms that the EPA was considering adopting for the communication of uncertainty). Remarkably, given the expert sample, the minimum probability associated with *likely* spanned four orders of magnitude, the maximum associated with *not likely* spanned five orders of magnitude, and the ranges assigned by committee members to these opposing terms overlapped.

Linguistic probabilities also serve other communicative functions that go beyond the vague probability levels they convey. This happens in part due to the *directionality* of probability phrases. If a term has negative directionality (e.g., *unlikely*), then strengthening it will convey even a lower probability of occurrence (e.g., *very unlikely*). Conversely, a term with positive directionality (e.g., *likely*) will convey a higher probability of occurrence when strengthened (e.g., *very likely*). Notably, terms having comparable numerical probability equivalents can differ in directionality (e.g., *quite uncertain* [negative] vs. *some possibility* [positive]). Directionality conveys information to receivers about a speaker's recommendations, preferences, or beliefs, serving as a linguistic frame (Teigen & Brun, 1999, 2003). Although the multiple functions of linguistic probabilities can increase the efficiency of communication—for example, by simultaneously offering a probabilistic estimate and making an implicit recommendation—in the realm of intelligence analysis, these dual functions could also mask potentially leading assessments by analysts with preferred policy stances (Piercey, 2009). Such situations can undermine policy neutrality and jeopardize analytic integrity.

Although requiring analysts to issue estimative assessments with numerical probabilities would be seen as a bridge too far by most in the IC (Barnes, 2015; Weiss, 2008), some organizations have implemented communications standards that attempt to regulate the use and meaning of linguistic probabilities. Such standards vary greatly. The NIC standard, for instance, establishes an ordinal scale with seven levels using eight probability terms, but the standard does not assign numerical probability equivalents to the terms. In contrast, U.K. Defence Intelligence (DI) uses 10 terms that form an ordinal scale with six levels, and each has an associated numerical probability range (Ho et al., in press). The ranges are mutually exclusive, unlike in standards used by other organizations such as the Intergovernmental Panel on Climate Change (IPCC; see Budescu, Por, Broomell, & Smithson, 2014), but there are intentional gaps between adjacent probability ranges. An alternative communication standard that has been used in some parts of the Canadian IC consists of 20 terms grouped into nine levels, which have point numerical probabilities associated with each level (Barnes, 2015). Although the numerical probabilities are represented as a point, they are intended to represent approximate values appropriately qualified by *about* (see Barnes, 2015). Some standards are printed in intelligence reports for consumers to read (e.g., the NIC, and U.K. DI), whereas others (e.g., the Canadian standard) have been used internally to improve analytic rigor and to support the scoring of forecast accuracy.

However, even when uncertainty communication standards are shared with users, users' interpretations of the terms tended to be in low agreement (Budescu et al., 2014). Recently, Ho et al. (in press, Study 2) showed that the agreement rate with probability terms in the U.K. DI standard could be improved by means of behavioral research methods that optimized the numerical ranges associated with each probability term. The present research applies a similar approach to the communication standard that was used in generating the forecasts analyzed in Mandel and Barnes (2014), which has not been studied previously. As noted earlier, the primary aim of this investigation is to infer the possible effects of miscommunication due to the use of linguistic probabilities on forecast accuracy from a consumer's perspective. This is an important step in forecast skill verification for systems that communicate probabilistic estimates to end users using words. It is also a requirement for verifying the quality of the communication standard. As Barnes (2015) described, the development of the standard was informed by a review of the published and unpublished literature on the interpretation of linguistic probabilities. Nevertheless, the standard is largely a product of one person's informed judgment. Thus, it remains to be seen how well subjects' interpretations of the terms conform to the intended meanings.

The present research also examined the effect of one aspect of context on the interpretation of linguistic probabilities. Studies have shown a variety of context effects on interpretation of linguistic probabilities. For instance, interindividual variability of interpretations is greater when a context is provided compared with when it is not (Beyth-Marom, 1982; Brun & Teigen, 1988). Translation of terms to numbers is also affected by prior beliefs about event base rates (Wallsten, Fillenbaum, & Cox, 1986; Weber & Hilton, 1990) and outcome severities (Harris & Corner, 2011; Weber & Hilton, 1990). Outcome valence has also been shown to influence the interpretation of linguistic probabilities. In one study (Mullet & Rivet, 1991), French primary and secondary school students assigned higher scaled probabilities to linguistic terms when they described an academic success than when they described an academic failure. However, subjects were not asked to translate linguistic probabilities to a numerical probability scale. Rather, they responded on a continuous 200 mm scale with anchors of *will certainly pass* and *will certainly fail*. The present experiment also examined each probability term in Barnes's (2015) communication standard within success and failure contexts. However, unlike Mullet and Rivet (1991), the interactive effect of probability term and context on *numerical* probability translations was also examined.

Experiment

Method

For obvious reasons, the policymakers who were the ultimate end users of the strategic intelligence forecasts studied in Mandel and Barnes (2014) were not accessible as subjects. Instead, a sample of 17 male military intelligence analysts in training at the Canadian Forces School for Military Intelligence and a sample of 21 female and 19 male University of Guelph undergraduates were studied. The analysts completed the task individually as part of an in-class exercise administered by the course instructor and were not paid for their participation. The undergraduates were recruited campus-wide and received a nominal fee for completing the task online.

A 20 (probability term) × 2 (outcome) within-subjects experimental design was used. The 20 probability terms used in the communication standard (Barnes, 2015) are shown in Table 1 and constitute the first factor of the design. The second factor refers to whether the term was used to describe an outcome that was either a success or a failure. The terms were embedded in a common sentence structure referring to an undisclosed operation that was either to fail or succeed. For instance, for the term *likely*, the sentence was "Operation X is likely to [fail/succeed]." Thus, a combination of 20 terms and 2 outcomes meant that subjects responded to 40 statements. The success and failure statements were run in blocks, and the order of the blocks was counterbalanced across subjects. Within each block, the order of statements was jumbled and is shown in the second column of Table 1. The order was jumbled to provide a more conservative test of subjects' interpretations, as phrase interpretations vary more

Table 1. Linguistic Probabilities on Percentage Chance Scale and Their Interpreted Numerical Meaning.

		Interpretation		
Probability Term	Order	Standard	Median	95% CI
Will not	20	0	0	[0, 0]
No prospect	7	0	2	[0, 05]
Little prospect	15	10	14	[11, 15]
Extremely unlikely	2	10	7	[5, 10]
Highly unlikely	18	10	10	[8, 10]
Very unlikely	19	10	10	[10, 12]
Low probability	3	20-30	25	[20, 25]
Probably not	5	20-30	25	[20, 27]
Unlikely	14	20-30	20	[20, 25]
Slightly less than even chance	16	40	45	[45, 45]
Even chance	9	50	50	[50, 50]
Slightly greater than even chance	1	60	55	[55, 55]
Probably	10	70-80	75	[70, 75]
Probable	12	70-80	71	[66, 75]
Likely	8	70-80	75	[75, 80]
Highly likely	11	90	85	[85, 90]
Extremely likely	6	90	90	[90, 90]
Almost certain	13	90	95	[90, 95]
Certain	17	100	100	[99, 100]
Will	4	100	100	[100, 100]

Note. CI = confidence interval.

across subjects when encountered in random order rather than as an ordered scale (Hamm, 1991). After each statement, subjects were asked to complete the sentence, "I interpret this statement to mean that the probability of [success/failure] is somewhere between ___% and ___%, with ___% as my best guess for a point estimate." For each blank, subjects were required to indicate a numerical value on a 0 to 100 percentage-chance scale.

Results

Given the aim of this research is to establish the most likely average interpretation for each linguistic probability in the communication standard, only data from subjects' best estimates are examined in this article. The data were statistically analyzed, and all ANOVA tests reported used the Greenhouse–Geisser correction for sphericity violations. A preliminary three-way ANOVA was run to examine whether sample (i.e., military analysts vs. students) as a dichotomous variable interacted with probability term or outcome to affect subjects' numerical translations. Neither the main effect of sample nor its two- or three-way interaction effects were statistically significant (i.e., *p*s > .05). Likewise, a three-way ANOVA with block order (i.e., failure statements presented before or after success statements), probability term, and outcome revealed no significant main or interaction effect of order on subjects' numerical translations. Therefore, subsequent analyses collapse over sample and block order.

Context effects on probability interpretation. All effects in the simplified two-way (Term × Outcome) ANOVA were significant: for term, $F(7.40, 414.45) = 413.86$, $p < .001$, $\eta_p^2 = .881$; for outcome, $F(1, 56) = 5.27$, $p = .026$, $\eta_p^2 = .086$; for Term × Outcome, $F(7.99, 447.54) = 3.85$, $p < .001$, $\eta_p^2 = .064$. As is to be expected, the main effect of probability term on numerical translations reflects subjects' discrimination of the terms across the standard. The main effect of outcome is due to subjects assigning a higher mean probability across terms in the failure condition ($M = 49.10$, 95% confidence interval [CI] = [47.69, 50.51]) than in the success condition ($M = 47.21$, 95% CI = [46.30, 48.12]). However, that effect was qualified by the interaction effect. Figure 1 plots estimated marginal mean numerical probability as a function of probability term and outcome. Two findings are noteworthy. First, the effect of context is stronger for linguistic probabilities ordered below *even chance* in the standard than for those ordered above *even chance*. Second, the terms are more effectively discriminated when they refer to probabilities of success than to probabilities of failure. In other words, the terms more effectively cover the full probability range from 0 to 1 when they refer to a success than when they refer to a failure.

Interpretation and usage of linguistic probabilities. An important aim of this research was to verify how well the prescribed meanings of the linguistic probabilities used in the communication standard described in Barnes (2015) agree with the

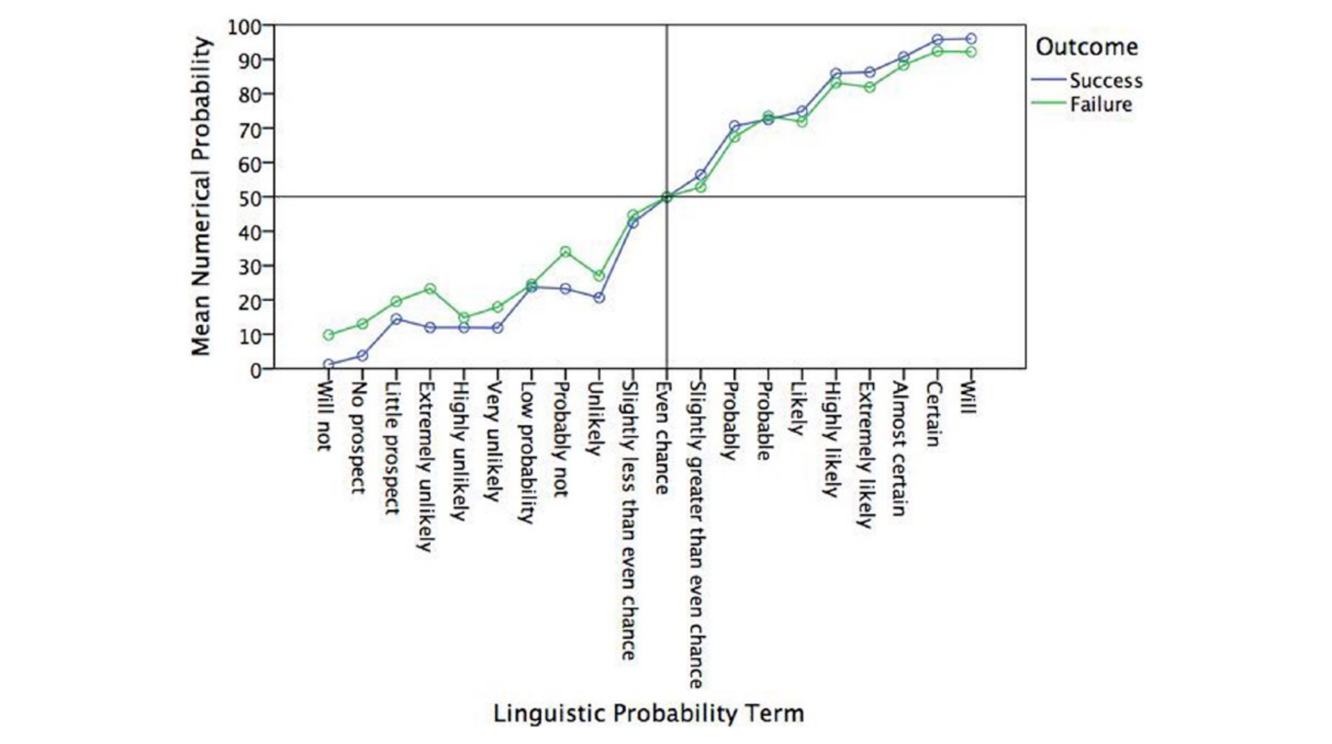

Figure 1. Estimated marginal mean numerical probability by probability term and outcome.

average interpretation of those terms by users (subjects, in this case). Column 3 in Table 1 shows the intended numerical equivalents for the 20 terms in the standard. Column 4 shows the median numerical probabilities subjects assigned as best point equivalents, and column 5 shows the 95% CIs around the median values that were generated using 1,000 bootstrap samples. There is strong agreement between the intended meaning of these terms and their average "received" meanings. Specifically, 17 out of 20 terms had intended meanings that fell within the 95% CI of the median estimate. Of the remaining three terms, *slightly less than even chance* and *slightly greater than even chance* were each 5 percentage points closer to the 50% midpoint than intended in the standard. The remaining term—*little prospect*—was assigned a 10% chance but might have been more appropriately assigned a 15% chance.

However, of the intelligence forecasts reported in Mandel and Barnes (2014) that associated numerical forecasts with terms from the standard (n = 1,396), the three terms just noted represented less than 4% of the forecasts. In fact, the first two terms were each used only once, and *little prospect* was used 33 times. Table 2 shows the frequency and percentage of term use in the forecast set. Although the term *very likely* was not included in the standard, it was used in 63 forecasts and is included in Table 2, and the associated forecasts are used in subsequent analyses. The term *would* was used in 25 forecasts and *would not* was also used 3 times. In Table 2 and subsequent analyses, these terms were merged with *will* and *will not*, respectively.

Table 2 also shows the directionality of each term. Of the 20 terms, 10 were positive, 9 were negative, and 1 (*even chance*) was neutral. Of the 1,391 forecasts (i.e., excluding the five forecasts that used *even chance*), 63.2% used terms with positive directionality and 36.8% used terms with negative directionality. The 99% CI around the former value is [59.8, 66.5]. Thus, a significant majority of forecasts were phrased using positive linguistic probability terms, even though the number of positive and negative terms in the standard was roughly balanced.

Accuracy of forecasts from the consumers' perspective. To assess the impact of communicating uncertainties in forecasts with linguistic probabilities, the median numerical probabilities shown in Table 2 (column 4) were substituted for the original forecast values used in Mandel and Barnes (2014). In that study, the mean Brier score was 0.074 based on 1,514 forecasts. The mean Brier score was recomputed for the subset of 1,369 forecasts for which median numerical estimates were available. The 63 forecasts using *very likely* were included and assigned the median value obtained for *highly likely*. The mean Brier score based on the original numerical probabilities assigned by analysts was 0.0724 (SD = 0.17), and it was 0.0711 (SD = 0.17) based on the median numerical probabilities obtained from the two participant samples. Given the large sample size, even this small mean difference was statistically significant, paired t(1368) = 2.21, p = .027, Cohen's d = 0.06. Yet, the effect size is very

Table 2. Frequency of Linguistic Probabilities in the Mandel and Barnes (2014) Forecast Set.

Probability Term	Frequency	% of 1,396	Direction	% Direction
Will/would	340	25.35	Positive	38.68
Likely	241	17.60	Positive	27.42
Unlikely	137	10.00	Negative	26.76
Very unlikely	117	8.55	Negative	22.85
Probably	95	6.94	Positive	10.81
Almost certain(ly)	90	6.57	Positive	10.24
Low probability	70	5.11	Negative	13.67
Very likely	63	4.6	Positive	7.17
Highly unlikely	38	2.78	Negative	7.42
Will/would not	35	2.56	Negative	6.84
Little prospect	33	2.41	Negative	6.45
Extremely unlikely	26	1.90	Negative	5.08
Probably not	26	1.90	Negative	5.08
Highly likely	26	1.90	Positive	2.96
Certain(ly)	20	1.46	Positive	2.28
Even chance	5	0.37	Neutral	100.00
Probable	2	0.15	Positive	0.23
No prospect	2	0.15	Negative	0.39
Slightly less than even chance	1	0.07	Negative	0.20
Slightly greater than even chance	1	0.07	Positive	0.11
Extremely likely	1	0.07	Positive	0.11

small. Using an alternative approach to forecast skill verification—area under the (relative operating characteristic) curve (AUC)—once again showed virtually no difference. In both cases, AUC = 0.94, the same value reported in Mandel and Barnes.

Discussion

This research verified how good the strategic intelligence forecasts studied by Mandel and Barnes (2014) were from the average intelligence consumer's perspective. Inferring that perspective from subjects' median "best" numerical translations of the linguistic probabilities, forecast accuracy was found to be as good as it was from the producer's perspective. By all quantitative measures reported, forecast accuracy could only be described as very good. It is also evident that the comparable forecast accuracy levels were due to the fidelity of the communication standard employed. The standard's prescribed numerical equivalents were well aligned with subjects' median estimates for most of the terms examined. This provides an important independent test of the standard's communication effectiveness.

The example of policy development and testing examined in this article also shows two distinct ways in which judgment and decision research can contribute to IC policy improvement. The first is by exploiting available research. As noted earlier, the communication standard was developed after extensively reviewing the existing literature on linguistic probability interpretation (Barnes, 2015). That knowledge exploitation in the development phase of the standard seems to have paid off given the high fidelity of the standard shown in this research. Second, applied behavioral science can play an important role in testing how well implemented policies, standards, and procedures are doing. Unfortunately, neither of these routes is currently common practice in the IC (National Research Council, 2011; Pool, 2010). The first is more frequent than the second, but the second is exceedingly rare. Yet, verification testing of proposed or already implemented policies or practices may be more important than knowledge exploitation because good ideas can also be developed in the absence of consulting the literature. Testing the efficacy of implemented policies, however, is a necessary requirement for effective organizational learning and for discriminating good ideas that succeed from good ideas that nevertheless fail.

The present findings also indicate how the communication standard examined here might be improved. Many terms were used infrequently, and all of the terms that had relatively weak agreement with the median estimates were ones that were used infrequently by analysts. Thus, the standard might be improved by using an abridged lexicon that eliminates several of the infrequently used terms. Some terms, such as *even chance*, may be worth retaining even if they are low frequency terms because they effectively define important points on the probability scale, such as the midpoint. Another possible improvement would be to test terms that are absent from the standard but which may be viable candidates for inclusion in a revised scale. These include complements of existing terms such as *certainly not, almost certainly not, high probability*, and *very likely*.

The findings also showed a context effect on linguistic probability interpretation. In contrast to Mullet and Rivet (1991), subjects in the present research assigned higher numerical probabilities, on average across terms, to failures than to successes. The finding is consistent with research showing that subjects judge combinations of failure to be more probable than combinations of success even when they are mathematically equiprobable (Mandel, 2008). It is evident, however, that the linguistic terms had weaker discrimination in the failure condition than in the success condition, and this was most pronounced for the low probability terms. Therefore, the interaction effect is what is of particular interest here.

It is unclear why this aspect of context moderated discrimination. One possibility is that success is a default representation and that failure is subsequently represented as "not succeeding." Moreover, people have difficulty interpreting phrases with linguistic probabilities when both the probability terms and the outcomes include implicit or explicit negations (e.g., it is unlikely [implicit negation] that Operation X will not occur [explicit negation]; Smithson, Budescu, Broomell, & Por, 2012). This may explain why discrimination was particularly undermined by the failure outcome for low probability terms, many of which include negations. For instance, if failure tends to be thought of as "not success," a statement such as "Operation X will not fail" might be read as "Operation X will_not not_succeed." Clearly, that is confusing, and more confusing than either "Operation X will_not succeed" or even "Operation X will not_succeed." However, even if failure is regarded as an implicit negation, the statement, "Operation X will not fail," may strike a reader as a double negation. A preference for affirmative phrases might also explain the observed preference for probability terms that were directionally positive because those terms do not include implicit or explicit negations.

The fact that a simple manipulation of context could influence the interpretation of probability phrases has implications for attempts to legislate meaning through standards for communicating uncertainty. Such effects reveal that the meaning of linguistic probabilities can only be assigned in an approximate manner even if a term is explicitly linked to a point probability in a reference standard. Policymakers choosing to institutionalize such approaches should realize that the contexts in which such terms appear will shape consumers' interpretations of what they read. Yet, it is worth noting that aspects of linguistic context influence the interpretation of numerical quantifiers too. For instance, people use contextual information to resolve whether numerical quantifiers refer to lower bounds, upper bounds, or precise values (Mandel, 2015).

Closing Remarks

Intelligence analysis is an exercise in expert human judgment under high accountability pressure and conditions of uncertainty. As such, it has much to learn from behavioral decision theory and research, which addresses precisely that sort of exercise. In recent years, the IC has shown considerable interest in strengthening its engagement with the behavioral sciences. This is evident in the U.S. IC through the funding of behavioral science research aimed at improving the human capabilities of intelligence analysts by ODNI's Intelligence Advanced Research Projects Activity and through ODNI's willingness to solicit recommendations and advice from the behavioral sciences (e.g., National Research Council, 2011; Pool, 2010). These are revolutionary developments. For far too long, the IC relied almost exclusively on Heuer's (1999) book, *Psychology of Intelligence Analysis*, as the summary of what behavioral science had to offer. Heuer's book, to its credit, provided a much-needed source of knowledge integration from literature on thinking and judgment, yet it offered nothing in the way of evidence-based testing of what works and what does not in intelligence analysis.

Behavioral science theories and methods can profoundly alter the way the IC understands, conducts, and monitors the analytic enterprise. Its theories could help generate hypotheses about which processes might work better than others; its methods could help devise sound tests of those ideas. For that to happen, the IC's leadership must come to see behavioral science as the best possible method at its disposal for process selection and performance monitoring. The present research exemplifies the sort of pragmatic approach needed to effectively make that case. The areas of research should be relevant to IC stakeholders—verification of forecast accuracy and communication fidelity are ones with which the IC is currently grappling (Dhami et al., in press; National Research Council, 2011). A key question is how far the IC will be willing to go in using behavioral science to answer practical questions about what works and what does not in intelligence analysis. The answer will ultimately shape any prospect of a behavioral science of intelligence analysis.

Acknowledgment

The author wishes to thank Alan Barnes, David Budescu, and Mandeep Dhami for their valuable feedback on an earlier draft of this article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Funding for this research was provided by Defence Research and Development Canada Applied Research Program Project 15dm (Understanding and Augmenting Human Analytic Capabilities) and Defence Research and Development Canada's Joint Intelligence Collection and Capability Project.

References

Adams, B. A., Thomson, M., Derbentseva, N., & Mandel, D. R. (2012). *Capability challenges in the human domain for intelligence analysis: Report on community-wide discussions with Canadian intelligence professionals* (Contract Report 2011-182). Toronto, Ontario: Defence Research and Development Canada. Retrieved from http://pubs.drdc-rddc.gc.ca/PDFS/unc118/p536570_A1b.pdf

Barnes, A. (2015). Making intelligence analysis more intelligent: Using numeric probabilities. *Intelligence and National Security*. Advance online publication. doi:10.1080/02684527.2014.994955

Betts, R. K. (2007). *Enemies of intelligence: Knowledge and power in American national security*. New York, NY: Columbia University Press.

Beyth-Marom, R. (1982). How probable is probable? Numerical translation of verbal probability expressions. *Journal of Forecasting*, *1*, 257-269.

Brun, W., & Teigen, K. H. (1988). Verbal probabilities: Ambiguous, context-dependent, or both? *Organizational Behavior and Human Decision Processes*, *41*, 390-404.

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

Clarke, V. A., Ruffin, C. L., Hill, D. J., & Beamen, A. L. (1992). Ratings of orally presented verbal expressions of probability by a heterogeneous sample. *Journal of Applied Social Psychology*, *22*, 638-656.

Dhami, M. K., Mandel, D. R., Mellers, B. A., & Tetlock, P. E. (in press). Improving intelligence analysis with decision science. *Perspectives on Psychological Science*.

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

Fingar, T. (2011). Analysis in the U.S. intelligence community: Missions, masters and methods. In B. Fischhoff & C. Chauvin (Eds.), *Intelligence analysis: Behavioral and social scientific foundations* (pp. 3-27). Washington, DC: National Academies Press.

Friedman, J. A., & Zeckhauser, R. (2012). Assessing uncertainty in intelligence. *Intelligence and National Security*, *27*, 824-847.

Hamm, R. M. (1991). Selection of verbal probabilities: A solution for some problems of verbal probability expression. *Organizational Behavior and Human Decision Processes*, *48*, 193-223.

Harris, A. J. L., & Corner, A. (2011). Communicating environmental risks: Clarifying the severity effect in interpretations of verbal probability expressions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1571-1578.

Heuer, R. J., Jr. (1999). *Psychology of intelligence analysis*. Washington, DC: Central Intelligence Agency, Center for the Study of Intelligence.

Ho, E., Budescu, D. V., Dhami, M. K., & Mandel, D. R. (in press). Communicating uncertainty effectively: Lessons from global climate change and intelligence analysis domains. *Behavioral Science & Policy*.

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

Mandel, D. R. (2008). Violations of coherence in subjective probability: A representational and assessment processes account. *Cognition*, *106*, 130-156.

Mandel, D. R. (2014). How good are strategic intelligence forecasts? *Policy Options*, *35*, 67-69.

Mandel, D. R. (2015). Communicating numeric quantities in context: Implications for decision science and rationality claims. *Frontiers in Psychology, 6*, Article 537. doi:10.3389/fpsyg.2015.00537

Mandel, D. R., & Barnes, A. (2014). Accuracy of forecasts in strategic intelligence. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 10984-10989.

Mandel, D. R., Barnes, A., & Richards, K. (2014). *A quantitative assessment of the quality of strategic intelligence forecasts* (Technical Report No. 2013-036). Toronto, Ontario: Defence Research and Development Canada.

Morgan, M. G. (1998). Commentary: Uncertainty analysis in risk assessment. *Human and Ecological Risk Assessment*, *4*, 25-39.

Mullet, E., & Rivet, I. (1991). Comprehension of verbal probability expressions in children and adolescents. *Language & Communication*, *11*, 217-225.

Murphy, A. H., Lichtenstein, S., Fischhoff, B., & Winkler, R. L. (1980). Misinterpretation of precipitation probability forecasts. *Bulletin of the American Meteorological Society*, *6*, 695-701.

National Research Council. (2011). *Intelligence analysis for tomorrow: Advances from the behavioral and social sciences*. Washington, DC: National Academies Press.

Office of the Director of National Intelligence. (2015). *Intelligence community directive 203: Analytic standards*. Washington, DC: Author. Retrieved from http://fas.org/irp/dni/icd/icd-203.pdf

Piercey, M. D. (2009). Motivated reasoning and verbal vs. numerical probability assessment: Evidence from an accounting context. *Organizational Behavior and Human Decision Processes*, *108*, 330-341.

Pool, R. (2010). *Field evaluation in the intelligence and counterintelligence context: Workshop summary*. Washington, DC: National Academies Press.

Scowcroft, B. (2009). *The value of intelligence analysis*. Address to the National Research Council Workshop on Behavioral and Social Science Research to Improve Intelligence Analysis [Audio file]. Retrieved from http://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_071411.mp3

Smithson, M., Budescu, D. V., Broomell, S. B., & Por, H.-H. (2012). Never say "not": Impact of negative wording in probability phrases on imprecise probability judgments. *International Journal of Approximate Reasoning*, *53*, 1262-1270.

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*, 155-190.

Teigen, K. H., & Brun, W. (2003). Verbal probabilities: A question of frame? *Journal of Behavioral Decision Making*, *16*, 53-72.

Tetlock, P. E. (2005). *Expert political judgment: How good is it? How can we know?* Princeton, NJ: Princeton University Press.

Tetlock, P. E., & Mellers, B. A. (2011). Intelligent management of intelligence agencies: Beyond accountability ping-pong. *American Psychologist*, *66*, 542-554.

Wallsten, T. S., & Budescu, D. V. (1995). A review of human linguistic probability processing: General principles and empirical evidence. *The Knowledge Engineering Review*, *10*, 43-62.

Wallsten, T. S., Budescu, D. V., Zwick, R., & Kemp, S. M. (1993). Preferences and reasons for communicating probabilistic information in numeric or verbal terms. *Bulletin of the Psychonomic Society*, *31*, 135-138.

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*, 571-587.

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*, 781-789.

Weiss, C. (2008). Communicating uncertainty in intelligence and other professions. *International Journal of Intelligence and CounterIntelligence*, *21*, 57-85.

Wheaton, K. (2012). The revolution begins on page five: The changing nature of NIEs. *International Journal of Intelligence and CounterIntelligence*, *25*, 330-349.

Yaniv, I., Yates, J. F., & Smith, J. E. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin*, *110*, 611-617.