Defence Research and Development Canada | Recherche et développement pour la défense Canada

CAN UNCLASSIFIED

DRDC | RDDC
technologysciencetechnologie

Geopolitical Forecasting Skill in Strategic Intelligence

David R. Mandel
DRDC – Toronto Research Centre

Alan Barnes
Carleton University

Journal of Behavioral Decision Making
doi: 10.1002/bdm.2055.

Date of Publication from Ext Publisher: October 2017

Defence Research and Development Canada
External Literature (P)
DRDC-RDDC-2017-P091
November 2017

Canada

IMPORTANT INFORMATIVE STATEMENTS

Geopolitical Forecasting Skill in Strategic Intelligence

David R. Mandel and Alan Barnes

Abstract

Extending research by the authors on intelligence forecasting, the forecasting skill of 3,622 geopolitical forecasts extracted from strategic intelligence reports was examined. The codable subset of forecasts (*N* = 2,013) was expressed with verbal probabilities (e.g., *likely*) and translated to numeric probability equivalents. This subset showed very good calibration and discrimination, but also underconfidence. There was no support for the hypothesis that forecasting skill was good mainly due to the general ease of forecasting topics. First, forecasting skill was as good among authoritative key judgments as in the general set. Second, forecasts that were assigned high degrees of certainty, indicative of ease, ($P \leq 0.05$ or $P \geq 0.95$) did not discriminate as well as less certain forecasts ($0.05 < P < 0.95$), and these subsets did not differ in calibration. Sensitivity and benchmarking tests further revealed that if the 1,609 uncodable forecasts were all assigned forecast probabilities of .5 (i.e., if all followed a "cautious ignorance" rule), skill characteristics would still show a large effect size improvement over a variety of guesswork strategies. The findings support a cautiously optimistic assessment of forecasting skill in strategic intelligence and indicate that such skill is not primarily attributable to the selection of easy forecasting topics. However, the large proportion of uncodable cases suggests that intelligence forecasts could be improved by avoiding imprecise language that not only affects codability, but also, in all likelihood, the interpretability and indicative value of forecasts for intelligence consumers.

Keywords: forecasting, prediction, intelligence analysis, skill, judgment

Forecasting is a vital part of intelligence assessment that enables tactical, operational, and strategic planning and decision-making. According to Allied intelligence doctrine, "analysis does more than look at the current situation, it should be predictive and therefore should address what might happen next, based upon alternative assumptions regarding the actions and reactions of different actors (including the impact of any intervention)" (NATO Standardization Office, 2016, §3.38). In particular, geopolitical forecasting offers anticipatory intelligence to key decision-makers, such as state leaders and other senior policymakers (Clapper, 2014). Timely, relevant, and accurate geopolitical forecasts promote national interests by reducing the probability of strategic surprises, mitigating national security risks, and allowing policymakers to capitalize on opportunities in a dynamically changing and globalized world.

Although the intelligence community assigns great importance to the anticipatory or estimative functions of intelligence, the quality of geopolitical forecasting remains largely unverified because intelligence organizations do not routinely monitor forecasting skill (Betts, 2007, Dhami, Mandel, Mellers, & Tetlock, 2015; Friedman & Zeckhauser, 2016). The decision not to verify forecasting skill on an ongoing basis with well-established scoring rules can have many deleterious consequences. First, without credible verification processes, intelligence organizations simply cannot know how good their forecasts are and how they might be improved. In the absence of such knowledge and subsequent corrective actions that might have been implemented, intelligence organizations increase their risk of failure to detect strategic threats and opportunities. Without proper verification, the intelligence community also remains unnecessarily susceptible to reactive pressure for institutional change—or even erasure (e.g., Moynihan (1991)—after politicized intelligence failures (Johnson, 2007; Tetlock & Meller, 2011). Without verification, the intelligence community forfeits opportunities for forecast improvement through calibration feedback (Rieber, 2004) or recalibration techniques aimed at mitigating institutional biases (Mandel & Barnes, 2014). Nor is the intelligence community well poised to evaluate the effects of training or structured techniques to improve analytic qualities such as forecasting accuracy if the latter is not carefully measured (Chang, Berdini, Mandel, & Tetlock, in press; Chang & Tetlock, 2016; Mandel, 2015b; Mellers et al., 2014).

Obtaining credible data for forecast verification purposes is a key challenge in many domains in which forecasting is an important organizational function. Task effects on forecasting accuracy can swamp individual differences (e.g., Stewart, Roebber, & Bosart, 1997), underscoring the importance of considerations about ecological validity and external validity. Organizations tracking their forecast accuracy will want to ensure that research supporting that objective is ecologically valid, although they may be far less concerned about external validity. Ideally, the forecasts would be those made by real experts doing their routine jobs under typical conditions. An organization studied in this manner may do little to promote externally valid findings if it happens to be atypical in important respects, but the findings would nevertheless be ecologically valid and of potential interest to those in the organization wishing to monitor performance and exploit the results. However, as the aim of such exercises veers towards scientific explanation of expert forecasting in a particular domain or even in general, questions

about the generalizability of results will demand greater attention.

Little is known about geopolitical forecasting skill, especially by intelligence organizations. Tetlock's (2005) long-term study of close to 300 political experts found mediocre forecasting skill. In general, experts were overconfident, and even the best experts (the uncertainty-tolerant Berlinian foxes) performed substantially worse than the best statistical models. Remarkably, experts who forecasted on topics in their areas of expertise did no better than dilettantes—namely, experts who forecasted on topics outside their areas of expertise. However, many topics in Tetlock's study required long-range forecasts that would only be resolved in years. In contrast, most intelligence forecasts (excluding futures or foresight exercises) are short- to medium-range and resolve in less than 1 year. A geopolitical forecasting tournament sponsored by the US government's Intelligence Advanced Research Program Activity (IARPA) elicited forecasts that generally resolved in 1 year or less. The winners found that elite "superforecasters" could be cultivated using a combination of effective sampling, elicitation, training, and aggregation methods (Mellers et al., 2014, 2015; Tetlock & Gardner, 2015). These results suggest that intelligence forecasts, with their similar time horizon, may be closer to the skill level shown in the IARPA tournament than in Tetlock (2005). However, the many differences between a tournament-style forecasting competition and on-the-job forecasting by intelligence analysts make such an inference highly tenuous. For one thing, superforecasters, by definition, are selected on the criterion. Secondly, they differ from "normal" forecasters in other respects. For instance, superforecasters better discriminate the meaning of verbal probability terms, they are less susceptible to content effects on their interpretation of such terms, and are more coherent on other judgment tasks (Mellers, Baker, Chen, Mandel, & Tetlock, 2017).

Ecologically valid studies of geopolitical forecasting skill in intelligence organizations are exceedingly rare. In one example, Lehner, Michelson, Adelman, and Goodman (2012) examined 187 geopolitical forecasts taken from unclassified or declassified intelligence reports and found poor discrimination but fairly good calibration. However, the methods used cast significant doubt on the interpretability of the study's findings. For instance, descriptive statements (e.g., "Arab groups in Kirkuk continue to resist violently what they see as Kurdish encroachment" [p. 730]) were rewritten as forecasts (i.e., "Arab groups in Kirkuk will resist violently what they see as Kurdish encroachment in the January 2007 to July 2009 time frame" [p. 731]. Another critical limitation is that verbal probability qualifiers in the original assessments were omitted in redrafted forecasts.

Mandel and Barnes (2014) conducted a long-term study of geopolitical forecasting skill in strategic intelligence, which examined 1,514 forecasts extracted from a comprehensive review of 6 years of classified reports produced by the Middle East and Africa (MEA) division of the Intelligence Assessment Secretariat (IAS) in the Canadian government. As part of regular analytic practice in the division investigated, analysts recorded whether their assessments were forecasts or other types of judgment (Barnes, 2016). Analysts assigned numeric probabilities to forecasts. The numeric probabilities were mapped to verbal probability terms following the lexical standard described in (Barnes, 2016; Mandel, 2015a), and only the verbal

probabilities appeared in finished intelligence reports. For example, one forecast (edited to remove sensitive information) was "It is very unlikely [1/10] that either of these countries will make a strategic decision to launch an offensive war in the coming six months." The numeric probability in brackets (i.e., $P = 0.10$) would not have been printed in the final report, but were recorded only for auditing and research purposes.

Forecasting skill based on the numeric probabilities analysts assigned was very good. The mean Brier score, *B*, is a proper scoring rule equal to the mean squared deviation between probabilities assigned to forecasts and outcomes coded 0 for non-occurrence and 1 for occurrence, which ranges from 0-1 (zero equaling a perfect score) (Brier, 1950; Murphy, 1973). The value of *B* reported in Mandel and Barnes (2014) was .074. This value is substantially better than that exhibited by all groups of forecasters studied by Tetlock, Mellers and colleagues in the IARPA tournament, with the exception of superforecasters' forecasts that were made in the final week before topics resolved. Under those conditions—namely, the most elite forecasters (selected based on their prior-year Brier scores) forecasting on very short timeframes—Mellers et al. reported a mean Brier score of .07 using a 0-2 scale, which equals .035 on the comparable Brier scale used by Mandel and Barnes. However, even superforecasters' mean Brier score of .125 on the 0-1 Brier scale in the initial week of forecasting was substantially higher than the mean Brier score reported in Mandel and Barnes (2014).

Mandel (2015a) re-examined the skill of the forecasts reported in Mandel and Barnes (2014) from the perspective of intelligence consumers reading the reports. To do so, students and intelligence analysts (namely, the potential consumers of probabilistic assessments) were asked to provide their best numeric equivalent for each of the 20 probability terms in the lexical standard that had been used by the MEA division (Barnes, 2016). Participants' median estimates were then substituted for analysts' numeric probabilities and the mean Brier score was recomputed. Using these median inferred probabilities, forecasting skill was virtually identical to that reported in Mandel and Barnes (2014) because the standard for mapping words to numbers used in the MEA division was well matched to participants' median estimates. In other words, forecasting skill was roughly the same from the producers' and consumers' perspective because the lexicon for probability terms stipulated meanings for those terms that were very close to the averaged interpretation of consumers.

One aim of the present research was to diversify the sample of forecasts studied in Mandel and Barnes (2014, Mandel, 2015) in order to test the robustness of the earlier findings. The present sample includes a large subset of forecasts from the original report but also includes forecasts (a) from interdepartmental committee reports that were generated by teams led by an IAS analyst and (b) from IAS reports produced by divisions other than MEA. These sources were not included in the earlier reports. The use of non-MEA division forecasts constitutes an important test of the generalizability of findings in Mandel and Barnes (2014, Mandel, 2015). The MEA division used several procedures that are atypical, such as requiring analysts to identify which of their assessments were forecasts and to assign numeric probabilities to forecasts (Barnes, 2016). Other IAS divisions did not use that approach and are more

representative of the intelligence community, in general. Indeed, it may be that the MEA division's atypical procedures contributed to the high skill level we reported earlier. Alternatively, general environmental features of intelligence organizations, such as accountability to skeptical stakeholders or requirements to be explicit about judgments and supporting reasoning, might account for skill (Arkes & Kajdasz, 2011).

A second research aim was to demonstrate how intelligence organizations could quantitatively assess their forecasting skill even if they do not use numeric probabilities in forecasts. This demonstration is important because intelligence organizations seldom require forecasts to be made with numeric crispness. If numeric probabilities were required for verifying skill, such a process would currently be infeasible, and it would likely remain so for some time to come given the high degree of institutional resistance to the idea of communicating probabilistic estimates using numbers (Barnes, 2016; Chang et al., in press; Friedman & Zeckhauser, 2016; Marchio, 2014). To illustrate the low-cost feasibility of verification, verbal probabilities in the forecast dataset that corresponded to terms in the lexicon described in Barnes (2016) were replaced with the median numeric probabilities elicited in Mandel (2015a). Verbal probabilities that were not in the lexicon were assigned numeric equivalents on the basis of synonymy to (or, in a few cases, interpolation between) terms in Barnes (2016). The general approach could easily be adopted with other lexicons used by other intelligence organizations or, more generally, by any organization that chooses to communicate uncertainties with verbal probabilities. In such cases, however, care should be taken to test the communication fidelity of those lexicons prior to use for such purposes (e.g., Budescu, Por, Broomell, & Smithson, 2014; Ho, Budescu, Dhami, & Mandel, 2015).

A final and subsidiary aim was to test the hypothesis advanced by Tetlock and Mellers (2014) that the high degree of forecasting skill reported in Mandel and Barnes (2014) may have been due to the selection of easy geopolitical topics. We test this facile-forecasting hypothesis in two, albeit indirect, ways. First, we examined skill separately for forecasts that were either key judgments or not. Key judgments appear at the outset of intelligence reports to convey the most important, informative, and authoritative assessments for intelligence consumers to consider. As Gries (1990) noted, "key judgments of [National Intelligence Estimates] are among the few written intelligence assessments regularly read at the top of government" (p. 2). Key judgments of IAS intelligence reports hold a comparable status in the Canadian context. While it is possible that key judgments include easy forecasts, it is less plausible that key judgments are mainly comprised of easy judgments. If so, they would be largely uninformative to intelligence consumers. Thus, if the facile-forecasting hypothesis were correct, forecasting skill on key judgments should be poorer than on the non-key judgments. Second, given that easy forecasts tend to be stated with high degrees of certainty (Lehner et al., 2012), the facile-forecasting hypothesis predicts an effect of grouping forecasts by certainty. Specifically, if forecasting skill were due to easy topic selection, skill would be expected to be significantly more impressive among the near-certain forecast subsample than among the less-certain forecasts subsample. Accordingly, we tested this hypothesis.

Method

Forecast Data

Forecast data were extracted from strategic intelligence reports produced by Canadian government sources that provide the Privy Council Office and other senior government clients with original, policy-neutral assessments of foreign developments and trends that may affect Canadian interests. All forecasts were reports that were either solely produced by the IAS (i.e., IAS intelligence memoranda) or by a small team led by an IAS analyst (interdepartmental committee reports). In all cases, analysts receive feedback from managers and peers. Therefore, the forecasts are best viewed as organizational rather than individual products. The data for this study were extracted from classified reports. However, the reported data separated the forecasts from their geopolitical content, and the data are unclassified. Representative examples of forecasts (edited to remove sensitive information) include the following: "For these reasons we believe that [development X] is very unlikely for at least the next six months" and "Nonetheless, the agreement will almost certainly prevent an outbreak of significant violence during the remainder of [year]."

Data were clustered into two unit subsamples: (a) forecasts from the IAS Middle East and Africa Division (MEA) produced by at least 11 analysts and (b) forecasts from other IAS divisions (NON-MEA) produced by at least 20 analysts. These numbers represent lower bounds because principal IAS analysts who led the production of interdepartmental committee reports were not uniquely identified. Thus, some of those analysts might be unaccounted for by the 31 analysts identified in IAS reports. The MEA subsample overlaps with data analyzed in Mandel and Barnes (2014; Mandel, 2015a), albeit imperfectly because present inclusion criteria differ in two respects. First, Mandel and Barnes (2014) restricted forecasts to those with analyst-assigned numeric probabilities, whereas any forecast that had an inferable numeric probability was included in this research (see details below). Second, Mandel and Barnes (2014) included forecasts from March 2005 to December 2011, whereas this research excluded forecasts prior to November 2006 to improve comparison with data obtained from the new subsamples.

In total, 3,622 forecasts were extracted from available reports, and 73% (2,629) had events that could be unambiguously coded as either having occurred or having not occurred. Of the latter, 77% (2,013) had verbal probability terms for which a numeric probability could be confidently assigned. This sample constituted the forecast data set for this research. Of the 2,013 forecasts, 1,735 (86%) were in the MEA subsample and 278 (14%) were in the NON-MEA subsample. As well, 1,759 (87%) forecasts were from IAS intelligence memoranda and 254 (13%) forecasts were from interdepartmental committees. There were 541 (27%) key judgments. If a forecast representing a key judgment was subsequently restated in a report, only the first was entered into the database and marked as a key judgment.

Coding Procedures

Procedures for outcome coding are reported in Mandel and Barnes (2014). In brief, a subject

matter expert who had established a 90% agreement rate with an independent subject matter expert on a sample of forecasts coded the outcomes into one of five categories: (a) event occurred, (b) event did not occur, (c) event partially occurred, (d) event partially did not occur, and (e) event cannot be determined. However, only cases in the first two categories were examined.

For each forecast, the verbal probability term used in the relevant report was recorded in a database. For terms in the lexical standard for communicating uncertainty used in the IAS-MEA sample (Barnes, 2016), the median numeric probability estimates elicited in Mandel (2015a), were used as inferred numeric probability equivalents (Table 1). As noted earlier, Mandel (2015a) asked a sample of students and intelligence analysts to provide their best numeric probability equivalent for the each of the 20 verbal probability terms in the Barnes (2016) lexicon. The median probabilities in Table 1 are based on study. Ninety-five percent of the terms used in the present study were drawn from that lexical standard and mapped directly to the median values in Table 1. As Table 2 shows, the remaining 5% were mapped to the closest synonym (see Reference Term column) and assigned that synonym's median value or else were assigned interpolated values from lower- and upper-bounding terms (i.e., where two terms are listed in the Reference Term column). For *almost certainly not*, we took the complementary numerical probability equivalent of *almost certain* (1 – .95 = .05) and extremized slightly to compensate for the fact that low-probability terms (e.g., *unlikely*) tend to have more extreme numerical probability equivalents than their high-probability counterparts (e.g., *likely*) (Brun & Teigen, 1988; Budescu, Weinberg, & Wallsten, 1988; Mandel, 2015a).

Results

We begin by reporting a set of analyses that examines the generalizability, or external validity, of findings in the earlier reports of Mandel and Barnes (2014) and Mandel (2015a). Next, we conduct two tests of the facile-forecasting hypothesis (Tetlock & Mellers, 2014). We then report the results of various benchmarking tests aimed at gauging how good forecasting skill is in the present sample. Finally, we examine the sensitivity of results to the inclusion of murky cases of forecasting that were excluded from our primary analyses (i.e., those cases where outcomes could not be unambiguously coded or where forecasts used terms for which a numeric probability could not be confidently assigned). Throughout the Results, we report 95% confidence intervals (CI) in square brackets. Unless otherwise stated, CI and *p* values are estimated in SPSS 20 statistical software by bias-corrected and accelerated simple bootstrapping with 1,000 samples per procedure.

Generalizability

As noted earlier, an aim of this research was to examine the generalizability of findings in our earlier reports, which evaluated the forecasting skill of strategic forecasts from a single division that employed various atypical analytical processes. Recall that the mean Brier score, $B = .074$ in Mandel and Barnes (2014) based on analysts' numeric probabilities, and $B = .071$ in Mandel (2015a) using inferred numeric probabilities from median sample estimates of the best

numeric equivalents to the verbal probability terms. Mean Brier scores in the new (non-MEA) IAS subsample (*B* = .050 [.030, .074]) and the IAS-led team subsample (*B* = .074 [.032, .126]) did not significantly differ, mean Δ = -.024 [-.081, .023], *p* = .31. These subsamples were subsequently collapsed to form the non-MEA subsample. The MEA subsample (*B* = .059 [.052, .067]) was virtually indistinguishable from the non-MEA subsample (*B* = .058 [.037, .084]), mean Δ = .0001 [-.024, .022], *p* = .99.

Although the Brier score is a proper scoring rule for probabilistic forecasts (Brier, 1950; Murphy, 1973), it is influenced by outcome variance, which is not a skill measure. It also does not differentiate the distinct skill components of calibration and discrimination (Yaniv, Yates & Smith, 1991). The area under the receiver-operator characteristic (ROC) curve, *A*, provides a useful measure of discrimination skill (Swets, 1986), including for analytic forecast quality (McClelland, 2011). *A*, the proportion of the total area of the unit square defined by the two axes of the ROC curve, can range from 0.5 (the area covered by the 45° no-discrimination line) to 1.0 (perfect discrimination). In Mandel and Barnes (2014), *A* = .94. In this research, *A* = .96 (*SE* = .006). As with the mean Brier score, discrimination did not significantly differ by unit: *A* = .96 (*SE* = .006) for MEA forecasts and *A* = .93 (*SE* = .021) for NON-MEA forecasts, *Z* = 1.14, *p* = .25. *Z* scores for comparisons of *A* are computed as follows (see Pearce & Ferrier, 2000):

$$Z = \frac{A_1 - A_2}{\sqrt{SE_{A_1}^2 + SE_{A_2}^2}}. \qquad [1]$$

Figure 1 plots the ROC curves for MEA and NON-MEA forecasts, both of which show very good discrimination.

Calibration was measured in two ways. First, we computed what Yates (1990) refers to as calibration in the large, C_L, the deviation between the mean subjective probability across all forecasts, $\bar{s}$, and the base rate of event occurrence over all forecasts, $\bar{o}$:

$$C_L = \frac{1}{N}\sum_{i=1}^{N}(s - o). \qquad [2]$$

In this research C_L = –.002 [–.013, .009], a value not significantly different from zero (perfect calibration), *p* = .71. Calibration did not differ between the two subsamples defined by unit: for the MEA subsample, C_L = –.003 [–.014, .009] and for the non-MEA subsample, C_L = 0.001 [–.025, .029], mean Δ = –.004 [–.033, .026], *p* = .81. Thus, observed calibration, overall, did not differ significantly from perfect calibration, and two unit subsamples showed comparable gross calibration skill.

Following Mandel and Barnes (2014) and consistent with Cox (1958) and Budescu and Johnson (2011), our second calibration test used a generalized linear model (GLM) to generate model-

based calibration curves. Unlike traditional calibration curves, which plot the relative frequency of outcome occurrences as a function of the outcome's forecasted probability of occurrence, the y-axis in model-based curves represents the mean model-predicted probability of the outcome. Model-based tests of calibration are particularly advantageous in cases where there is sparse data in some of the forecast probability bins (Budescu & Johnson, 2011), as is the case in this study.

The GLM we used to test calibration had a binary logistic link function with forecast, unit, the two-way interaction with forecast as predictors and outcome as the criterion. As expected, forecast, which was treated as a continuous variable in the analysis, was a significant predictor of outcome, B = 7.07 [5.62, 8.53], Wald χ^2 = 90.70, p < .001. Unit was nonsignificant, B = –1.01 [–2.11, 0.09], Wald χ^2 = 3.25, p = .072. However, the forecast × unit interaction was significant, B = 1.71 [0.08, 3.33], Wald χ^2 = 4.25, p = .039. The mean predicted values of the model were used to plot the calibration curves shown in Figure 2. The spline-smoothed sigmoid calibration curves shown in Figure 2 reveal underconfidence in forecasting, consistent with that found in Mandel and Barnes (2014). From Figure 2, it appears that underconfidence is weaker in the NON-MEA subsample.

To directly test the difference in underconfidence between the MEA and NON-MEA subsamples, a calibration of confidence index (Lichtenstein & Fischhoff, 1977), C_C, was computed:

$$C_{Conf} = \frac{1}{N_1 + N_2}\left(\sum_{i=1}^{N_1}(o - s) \text{ iff } s < .5 + \sum_{i=1}^{N_2}(s - o) \text{ iff } s > .5\right). \quad [3]$$

Negative values of C_{Conf} indicate underconfidence, whereas positive values indicate overconfidence. In the MEA subsample, the effect of underconfidence was small to medium in size, C_{Conf} = –.084 [–.094, –.074], p < .001, Cohen's d = 0.37, whereas in the non-MEA subsample, the effect was very small and marginally significant, C_{Conf} = –.031 [–.059, .001], p = .052, Cohen's d = 0.13. As Figure 2 indicated, MEA forecasts were significantly more underconfident than NON-MEA forecasts, although the effect is small, mean Δ = –.053 [–.089, –.024], p = .001, Cohen's d = 0.23.

Difficulty

To test the facile-forecasting hypothesis (Tetlock & Mellers, 2014), we first compared forecasts from key judgments to forecasts not in key judgments. Because key judgments are the most authoritative and informative assessments made in intelligence reports, we reasoned that they are less likely than other forecasts to address easy topics, which would be uninformative to decision-makers. One indication that key judgments are more difficult is an outcome variance estimate closer to .25 (maximum uncertainty). Outcome variance, V, is a non-skill related component of the Brier score, is thus defined:

$$V = \bar{o}(1 - \bar{o}). \quad [4]$$

In Eq. 4, $\bar{o}$ is the base rate of event occurrence in a population or sample. The variance of outcomes is .245 in the key-judgment subsample and .230 in the non-key-judgment subsample. The base rates used to calculate these values differ significantly: $\bar{o}$ = .571 [.529, .613] in the key-judgment subsample and $\bar{o}$ = .640 [.615, .664] in the non-key judgment subsample, mean Δ = .069 [.021, .117], p = .005.

Contrary to the facile-forecasting hypothesis, discrimination skill in the key-judgment subsample was actually better (A = .967, SE = .008) than in the non-key-judgment subsample (A = .952, SE = .007), although the difference was nonsignificant, Z = 1.41, p > .15. Calibration did not significantly differ from 0 (i.e., perfect calibration) in either subsample (ps > .35): for key judgments, C_L = .008 [–.012, .029] and for non-key-judgments, C_L = –.006 [–.019, .008], and these values did not significantly differ from each other, mean Δ = –.013 [–.036, .008], p > .24.

A second test of the facile-forecasting hypothesis involved comparing the 40% of cases in which forecasts were assigned very high certainty (i.e., probabilities greater than or equal to .95 or less than or equal to .05) with the remaining 60% of cases that were less certain (i.e., probabilities greater than .05 but less than .95). As expected, outcome variance was substantially smaller in the high-certainty subsample (V = .12) than in the lower-certainty subsample (V = .25) in the lower-certainty subsample. Contrary to the facile-forecasting hypothesis, discrimination was significantly worse in the high-certainty subsample (A = .89, SE = .023) than in the lower-certainty subsample (A = .96, SE = .007), Z = –2.70, p = .007. Calibration also differed between the two subsamples: C_L = .017 [.003, 030] in the high-certainty subsample and C_L = –.015 [–.030, –.000] in the lower-certainty subsample, mean Δ = .031 [.012, .052], p = .005. However, it is evident that this effect is due to the differing signs of the calibration values rather than their magnitudes. The absolute difference in calibration between the two subsamples is negligible, mean Δ = .002 [–.016, .021], p = .85.

Benchmarks

As noted earlier, it is difficult to compare skill scores across forecasting studies because the studies differ on multiple dimensions that can affect the expression of forecasting skill. To gauge “how good” forecasts are, it is thus useful to generate benchmarks within the same study. For instance, forecasting skill by intelligence organizations could be compared to various baseline measures. One measure—the equivalent of a dart-throwing chimp, as Tetlock (2005) put it—is to randomly guess whether the event will occur or not and to express the forecasts with complete certainty (i.e., randomly assigning probabilities of 0 or 1). A slightly more sophisticated rule could take account of the base rate of event occurrences. In this research, the base-rate tracking rule would involve random forecasting with the constraint that 62% of forecasted outcomes had to be occurrences (namely, because $\bar{o}$ = .62). A third rule is to follow the base-rate rule but assign less certainty to forecasts. To implement the “wary-base-rate” rule, we used probabilities of .25 for expected non-occurrences and .75 for expected

occurrences. For the first rule, we generated 10 random sequences of 2,013 zeros and ones.[1] For the second rule, we again generated 10 sequences but they were drawn randomly from a set of 62 ones and 38 zeros, thus ensuring that $E(\bar{o}) = .62$. For the third rule, we recoded all zeros and ones in the second rule as .25 and .75, respectively.

Brier scores were computed for each of the 30 sequences, and the mean Brier score for each of the 10 sequences within each rule type was calculated. The grand means (i.e., the mean over items and randomization runs) for the chimp, base-rate, and wary-base-rate rules were .501 [.495, .508], .474 [.466, .483], and .300 [.295, .304], respectively. These values differ significantly from each other (ps < .001 by bootstrap paired t tests). Clearly, all are significantly worse than the mean Brier score based on analytic forecasts (B = .059 [.052, .066]). Indeed, since the three rules perform worse than the "cautious ignorance" rule of invariably forecasting .5 for all topics (which yields B = .25), a better benchmark test is to compare the mean Brier score to the value of .25. A one-sample t test comparing B to .25 yields t = 55.12. The resulting effect size is very large, Cohen's d = 1.23. At the other end of the benchmarking spectrum, we could compare the observed Brier scores to perfect scoring (B = 0). A one-sample t returns a value of 16.84, constituting a small-to-medium effect size, Cohen's d = 0.38. Finally, if we assign forecasts of .5 to the 1,609 forecasts that were excluded from the primary analyses and include them with our primary sample, comparable analyses with the resulting Brier score of .144 would yield large effect sizes for comparisons to B = .25 (Cohen's d = 0.71) and B = 0 (Cohen's d = 0.96).

Sensitivity

The high skill level observed in the forecast data naturally raises questions about the extent to which such performance is due to biased sampling. As noted earlier, there were 1,609 forecasts (44.4%) excluded either because the outcome could not be coded as a clear case of event occurrence or non-occurrence or because the linguistic terms used to describe the likelihood of the event (e.g., *could* or *might*) were not translatable into numeric probabilities. A conservative test of the effect of data exclusion would be to treat each of these excluded cases as a forecast of maximum uncertainty (s = .5), reflecting a "cautious ignorance" rule. In this case, the Brier score would equal .25 whether or not the outcome occurred and, as noted in the preceding section, the mean Brier score for the full sample of 3,622 cases would equal .144. Moreover, if it were assumed that $\bar{o} \approx .62$ among the excluded cases, as it equals in the primary sample, then $C_L \approx .06$. Unsurprisingly, these adjusted statistics reflect significantly poorer skill than those based on the primary dataset.

Finally, we tested the sensitivity of the mean Brier score to the number of forecasting categories used in this research (16 levels of probability) by reassigning forecast values to their nearest value on the original 9-point scale used in Mandel and Barnes (2014). For values of .05 and .95, which are equidistant to the lower and upper adjacent values on the 9-point scale, we calculated the Brier score based on rounding upwards and on rounding downwards and then took the average. Based on the 9-valued scale, B = .060 [.054, .067], which represents a small but significant increase over the mean Brier score based on the 16-valued scale, mean Δ =

.0019 [.0014, .0023], $p < .001$, Cohen's $d = 0.18$. However, we also remapped forecasts to the 7-valued "recalibration" scale that improved calibration in Mandel and Barnes (2014). This significantly improved (i.e., lowered) the Brier score ($B = .045$ [.037, .053]) over that based on the 16-valued scale, mean $\Delta = -.013$ [–.015, –.011], $p < .001$, Cohen's $d = 0.27$.

Discussion

In our earlier report on the quality of strategic intelligence forecasts (Mandel & Barnes, 2014), we concluded that the findings warranted tempered optimism. The present findings continue to warrant tempered optimism about the quality of geopolitical forecasting by intelligence organizations. Clearly, no single study on this topic is definitive. There is much unknown about the parameters that may affect forecasting skill, in general, and in the context of intelligence production, in particular. Thus, a primary aim of this research that was prompted by our earlier report on accuracy of strategic intelligence forecasts was to examine the generalizability of those findings. We addressed that issue by examining forecasts from other divisions of the same organization that produced the forecasts assessed in Mandel and Barnes (2014) and by small inter-organizational teams of analysts.

Although the additional samples were smaller than the original, the results unambiguously support the generalizability of the earlier findings: in both MEA and NON-MEA subsamples, discrimination and calibration skill were comparable. The largest observed difference between these subsamples was the degree of underconfidence exhibited, but there too the similarities outweighed the differences. The effect size was small and both subsamples showed underconfidence. The convergence of findings across the subsamples examined in this research suggest that forecasting skill may be a more general characteristic of strategic intelligence assessment and not linked to the idiosyncratic analytic procedures described in Barnes (2016), such as use of numeric probabilities in forecasting. A reasonable conjecture is that the analytic standards introduced in the MEA division prompted analysts to be more cautious in their assessments, and that might explain the greater degree of underconfidence exhibited in the MEA subsample than in the non-MEA subsample.

A second aim of our study was to show how researchers and quality control evaluators within intelligence organizations (or other organizations that use verbal probabilities to communicate forecasts or other probabilistic judgments) could still apply quantitative scoring rules that enable judgment skill verification. Although numeric probabilities may be beneficial for communicating uncertainty, this research shows that it is unnecessary to have numeric estimates provided by experts in order to measure facets of their forecasting skill. Recent studies (Ho et al., 2015; Mandel, 2015a) have examined how probability terms used in the intelligence communities of various countries are interpreted, and such information can be used to infer numeric probabilities where only verbal probabilities have been assigned. This is an important methodological fact because quantification of uncertainty in intelligence is rare and there remains opposition to quantifying probabilities and expressions of uncertainty (Friedman & Zeckhauser, 2016; Marchio, 2014; Spielmann, 2016), much as Kent (1964) noted over a half-century ago, and as has been observed across a wide range of expert communities

(Morgan, 2014). Intelligence organizations should not have to wait for acceptance and use of numeric probabilities in intelligence analysis before they can proactively verify forecasting skill components, and this research demonstrates that they don't have to wait.

A third aim of this research was to test Tetlock and Mellers' (2014) facile-forecasting hypothesis that good forecasting skill is simply due to topics that are easy to predict. If the facile-forecasting hypothesis were correct, it would call into question why governments spend vast sums of money tackling prediction problems that are easily or already known. Thus, it is important to probe how widespread easy forecasting is within the intelligence community. Unfortunately, direct assessments of difficulty are hard to make in field studies where content is often classified and stimuli cannot be pretested to find probable solution rates for various populations. Nevertheless, indirect tests can be conducted and their results can be triangulated. The process is admittedly far from epistemologically ideal, but we believe it has merit as long as the interpretation of findings remains appropriately skeptical.

The indirect tests we conducted yielded consistent findings that cast doubt on the facile-forecasting hypothesis. First, as already noted, the findings generalize the results of our earlier work (Mandel & Barnes, 2014; Mandel, 2015a). Generalizable skill ramps up the charge against the intelligence community posed by the facile-forecasting hypothesis, and it correspondingly weakens the plausibility of the hypothesis by raising the burden of proof. Second, we found that forecasts that constituted key judgments were no less skillful than other less illustrious forecasts made in intelligence reports. Key judgments reflect the intelligence producer's assessment of their best hand—the most important judgments to communicate to decision-makers who may not have time to read full reports. It is reasonable to assume that the most informative judgments produced by an intelligence unit are at least positively correlated with task difficulty. Thus, contrary to the findings, the facile-forecasting hypothesis predicts that key judgments would decline in skill sharply. Finally, the facile-forecasting hypothesis predicts a large percentage—perhaps even a majority—of forecasts would be made with high certainty (due to their ease), and that subset should appear much more skillful than lower-certainty forecasts. Yet contrary to that prediction, partitioning forecasts by uncertainty level had no effect on calibration and the opposite effect on discrimination. That is, discrimination was not as good in the high-certainty subsample. Taken together, our findings cast doubt on the veracity of the facile-forecasting hypothesis.

Prescriptive Implications

Our research highlights areas where intelligence forecasting could be improved. Over a quarter (27%) of forecasts extracted from intelligence products were not clear enough to be coded as occurrences or non-occurrences with sufficient confidence, thus failing Tetlock's (2005) clairvoyance test—namely, that a clairvoyant would be able to answer a forecasting question with complete confidence and accuracy because vagueness and ambiguity had been exorcised from the question. Forecasts failing the clairvoyance test are arguably of diminished value to decision makers for a variety of reasons: they make misinterpretations of estimates more likely, they do less than unambiguous estimates to reduce uncertainty, and they only weakly

bolster decision-makers' accountability. Intelligence organizations should implement procedures that help analysts and their managers express forecasts clearly. For instance, in the process of reviewing intelligence forecasts, analysts and their managers could consider the range of possible outcomes and judge whether the forecast, as stated, sufficiently expresses the analytic assessment of the different probabilities of possible outcomes. If the answer is murky to the assessment producers, it will likely be murkier to consumers (or worse—ostensibly clear but, in fact, misleading).

We also found that a substantial proportion (23%) of forecasts with codable outcomes used linguistic terms to convey uncertainty that could not be mapped onto numeric equivalents with adequate confidence. Attempts to minimize the usage of "weasel words" and other terms that detract from clear interpretation go back over half a century in intelligence (e.g., Kent, 1964), and our findings indicate that the need to address the issue lingers on. The use of evidence-based standards for communicating uncertainty might help by increasing the correspondence between stipulated and personally held interpretations of verbal probability terms (e.g., Barnes, 2016; Ho et al., 2015). Such standards could be supplemented with a product review process that promotes and verifies compliance with the standards.

A more radical step would be to use numeric probabilities in intelligence assessments. Decision science could contribute much to explaining why doing so would be beneficial, why it would not be inappropriate, and how it could be done in ways that take the intelligence community's values into consideration. For instance, intelligence organizations cling to verbal probabilities in part because of the blame deflecting wiggle room they apparently provide in comparison to numbers. In this regard, intelligence organizations are like most individuals who prefer to communicate uncertainties verbally but receive them numerically (Brun & Teigen, 1988; Wallsten, Budescu, Zwick, & Kemp, 1993). Yet, verbal probabilities may be more blame inviting than deflecting. Jenkins, Harris, and Lark (2017) gave participants probabilistic predictions in numerical (e.g., "20% likelihood"), numerical-verbal (e.g., "20% likelihood [unlikely]"), verbal-numerical (e.g., "unlikely [20% likelihood]"), or verbal (e.g., "unlikely") formats, and then informed them that the prediction was erroneous. Numeric estimates were viewed as least incorrect and verbal estimates as most incorrect, and assessments of the forecaster's credibility showed a similar pattern. More importantly, the precision enabled by numeric estimates has recently been demonstrated to improve geopolitical forecast accuracy (Friedman, Baker, Mellers, Tetlock, & Zeckhauser, in press).

Limitations and Future Directions

Notwithstanding the informative value of this study, we urge caution in interpreting its results. As already noted, many forecasts lacked sufficient clarity in event and uncertainty descriptions to be subjected to scoring rules. Sensitivity testing permitted an estimate of the expected lower bound on forecasting skill if those forecasts were included (i.e., a mean Brier score equal to .14). That worst-case estimate still showed a large skill advantage over the cautious ignorance rule—effectively, the best one could hope to eke out of a no-skill rule. Yet it would be misleading to ignore cases that could not be scored, and focus only on the very good

performance in the set that was scored.

We also urge caution in interpreting results given that the distribution of forecasts over the probability scale afforded relatively few cases to judge performance in the mid-range of the scale. In fact, only 9/2,013 forecasts—less than half a percent—were in the middle half of the probability scale (i.e., greater than .25 and less than .75). Despite the large sample size, estimates of forecasting skill are computed over probability ranges with sparse data. This characteristic, however, is unlikely to represent an anomaly of our sample. Because fence-sitting forecasts close to fifty-fifty are unlikely to be very informative, analysts may try or may be directed to try to avoid them (Barnes, 2016). In this respect, the sparse data in the midrange of the probability scale is likely to reflect the representative design of this research rather than a sampling failure.

Our findings also underscore the need for widespread verification of forecasting skill in intelligence organizations so that the parameters that affect skill can be identified and estimated. To date, investigation has focused mainly on Canadian strategic intelligence forecasts made by civilian intelligence analysts. Hypotheses could be advanced about each of these attributes as putative moderators of forecasting skill. For instance, studies of strategic intelligence forecasting may tell us little about the skill of tactical or operational forecasters. Likewise, because the US is a superpower that usually leads Canada in decisions about military intervention and other foreign policy matters, US strategic forecasting may be more challenging and might show different skill characteristics. Studies of the kind reported here and in Mandel and Barnes (2014) should be replicated in other allied countries, across a wide range of intelligence organizations, and covering distinct types of forecasting requirements.

Where feasible, the findings of such research should be triangulated with evidence from classified or declassified reports probing forecast content, which would provide a more detailed picture to the intelligence community. The CIA's trove of declassified validity studies could be examined for such purposes (Marchio, 2016). These and other triangulation attempts are vital for a comprehensive understanding of forecasting quality in intelligence because research such as that reported here and in our earlier reports focuses only on forecasts that were made. The point also applies to IARPA's recent geopolitical forecasting tournament (Mellers et al., 2014; Tetlock & Gardner, 2015). Such studies, while highly informative, do not inform us about how well intelligence organizations detect low frequency, high severity, "black swans" that typically fall off the radar (Makridakis & Taleb, 2009)—forecasts that were not made but that, in hindsight, clearly ought to have been. In tournaments where forecasting topics are cultivated by a research team, the prospect of studying the process of forecast topic generation is simply precluded. Yet, Mandel, Barnes, and Richards (2014) found that 70% of strategic intelligence forecasts were on topics developed "in house" and 30% were on client-driven topics, with the latter showing better skill characteristics. How the topic generation process unfolds "in the wild" is a topic that could be profitably pursued in future research.

Conclusion

The present research provided a vital test of the generalizability of a rare naturalistic study of geopolitical forecasting skill in strategic intelligence. The test was vital because the earlier study focused on an organizational unit that had employed various atypical analytic practices. The present research suggests that those practices had little effect on forecasting skill characteristics, except perhaps to make forecasts even more underconfident. Our study, however, provides no clear answer to the question of why geopolitical forecasting skill in strategic intelligence is as good as it appears to be.

We maintain, as we had in our earlier report, that a key characteristic of forecasting in intelligence is that the forecasters (i.e., intelligence analysts) are decoupled from decision-making and are thus primarily accountable for the quality of their judgments. Accountability pressure can have salutary effects on judgment, such as reducing overconfidence (Tetlock & Kim, 1987), prompting deeper information processing (Chaiken, 1980) and heightening awareness of the informational determinants of personal choices (Hagafors & Brehmer, 1983). Accountability pressures on analysts are far greater than those placed on forecasters in geopolitical tournaments (Arkes & Kajdasz, 2011). As advisors on matters of national security to their nations' leaders and policymakers, they face the ultimate in skeptical audiences having to, as the cliché goes, "speak truth to power" on a routine basis. We hypothesize that individuals drawn to intelligence analysis tend to exhibit a Berlinian fox-like disposition (i.e., tolerant of uncertainty and cognitive conflict), which predicts good forecasting (Tetlock, 2005). We hypothesize further that a fox-like disposition is reinforced among analysts by the accountability pressures and skeptical culture they encounter on the job. These hypotheses suggest an interactionist account of forecasting skill in intelligence that deserves rigorous testing in future research.

References

Arkes, H. R., & Kajdasz, J. (2011). Intuitive theories of behavior. In B. Fischhoff & C. Chauvin (Eds.), *Intelligence Analysis: Behavioral and Social Scientific Foundations* (pp. 143-168). Washington, DC: National Academies Press.

Barnes, A. (2016) Making intelligence analysis more intelligent: Using numeric probabilities. *Intelligence and National Security, 31*, 327-344.

Betts, R. K. (2007). *Enemies of Intelligence: Knowledge and Power in American National Security.* New York: Columbia University Press.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*, 1-3.

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., & Johnson, T. R. (2011). A model-based approach for the analysis of the calibration of probability judgments. *Judgment and Decision Making, 6,* 857–869.

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

Budescu, D. V., Weinberg, S., & Wallsten, T. S. (1988). Decisions based on numerically and verbally expressed uncertainties. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 281-294.

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message and cues in persuasion. *Journal of Personality and Social Psycholology*, *39*, 752-766.

Chang ,W., Berdini, E., Mandel, D. R., & Tetlock, P. E. (in press). Restructuring structured analytic techniques in intelligence. *Intelligence and National Security.*

Chang ,W., & Tetlock, P. E. (2016). Rethinking the training of intelligence analysts. *Intelligence and National Security, 31,* 903-920.

Clapper, J. (2014). *The National Intelligence Strategy of the United States of America: 2014*. Washington, DC: Office of the Director of National Intelligence.

Cox, D. R. (1958). Two further applications of a model for binary regression. *Biometrika*, *45,* 562–565.

Dhami, M. K., Mandel, D. R., Mellers, B. A., & Tetlock, P. E. (2015). Improving intelligence analysis with decision science. *Perspectives on Psychological Science, 106*, 753-757.

Friedman, J. A., Baker, J. D., Mellers, B. A., Tetlock, P. E., & Zeckhauser, R. (in press). The value of precision in probability assessment: Evidence from a large-scale geopolitical forecasting tournament. *International Studies Quarterly.*

Friedman, J. A., & Zeckhauser, R. (2016). Why assessing estimative accuracy is feasible and desirable. *Intelligence and National Security, 31*, 178-200.

Gries, D. D. (1990). New links between intelligence and policy. *Studies in Intelligence, 34*, 1-7.

Hagafors, R., & Brehmer, B. (1983). Does having to justify one's decisions change the nature of the judgment process? *Organizational Behavior and Human Performance*, *31*, 223-232.

Ho, E., Budescu, D. V., Dhami, M. K., & Mandel, D. R. (2015). Improving the communication of uncertainty in climate science and intelligence analysis. *Behavioral Science & Policy* 1, 43-55.

Jenkins, S. C., Harris, A. J. L., & Lark, R. M. (2017). Maintaining credibility when communicating uncertainty: the role of communication format. In G. Gunzelmann, A.
Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 582-587). Austin, TX: Cognitive Science Society.

Johnson, L. K. (2007). A shock theory of congressional accountability for intelligence. In L. K. Johnson (Ed.), *Handbook of Intelligence Studies* (pp. 343-360). New York: Routledge.

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

Lehner, P., Michelson, A., Adelman, L., & Goodman, A. (2012). Using inferred probabilities to measure the accuracy of imprecise forecasts. *Judgment and Decision Making, 7*, 728–740.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior & Human Performance, 20,* 159-183.

Makridakis, N., & Taleb, N. (2009). Living in a world of low levels of predictability. *International Journal of Forecasting, 25*, 840-844.

Mandel, D. R. (2015a). Accuracy of intelligence forecasts from the intelligence consumer's perspective. *Policy Insights from the Behavioral and Brain Sciences, 2*, 111-120.

Mandel, D. R. (2015b). Instruction in information structuring improves Bayesian judgment in intelligence analysts. *Frontiers in Psychology, 6*, article no. 387, doi:10/3389/ fpsyg.2015.00387

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 2013-036]. Toronto, Canada: Defence Research and Development Canada.

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.

Marchio, J. (2016). "How good is your batting average?" Early IC efforts to assess the accuracy of estimates. *Studies in Intelligence, 60*(4), 3-13.

McClelland, G. H. (2011). Use of signal detection theory as a tool for enhancing performance and evaluating tradecraft in intelligence analysis. In B. Fischhoff & C. Chauvin (Eds.), *Intelligence Analysis: Behavioral and Social Scientific Foundations* (pp. 83-99). Washington, DC: National Academies Press.

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*, 369-381.

Mellers, B. A., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., Murray, T., Stone, E. & Tetlock, P. E. (2014). Psychological strategies for winning geopolitical forecasting tournaments. *Psychological Science, 25*, 1106–1115.

Mellers, B. A., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L., & Tetlock, P. E. (2015). Identifying and cultivating "Superforecasters" as a method of improving probabilistic predictions. *Perspectives in Psychological Science, 10*, 267–281.

Morgan, M. G. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National Academy of Sciences of the United States of America, 111,* 7176-7184.

Moynihan, D. P. (1991, May 19). Do we still need the C.I.A.? The State Dept. can do the job. *New York Times*, E17.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology, 12*, 595–600.

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

Rieber, S. (2004). Intelligence analysis and judgmental calibration. *International Journal of Intelligence and Counter-Intelligence, 17*, 97–112.

Spielmann, K. (2016). I got algorithm: Can there be a Nate Silver in intelligence? *International Journal of Intelligence and CounterIntelligence, 29*, 525-544

Stewart, T. R., Roebber, P. J., & Bosart, L. F. (1997). The importance of the task in analyzing expert judgment. *Organizational Behavioral and Human Decision Processes, 69,* 205-219.

Swets, J. A. (1986). Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance. *Psychological Bulletin, 99*, 181-198.

Tetlock, P. E. (2005). *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton, NJ: Princeton University Press.

Tetlock, P. E., & Gardner, D. (2015). *Superforecasting: The Art and Science of Prediction* New York: Crown.

Tetlock, P. E., & Kim, J. I. (1987). Accountability and judgment processes in a personality prediction task. *Journal of Personality and Social Psychology*, *52*, 700-709.

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

Tetlock, P., & Mellers, B. (2014). Judging political judgment. *Proceedings of the National Academy of Sciences of the United States of America, 111*, 11574-11575.

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

Yates, J. F. (1990). *Judgment and Decision Making.* Englewood Cliffs, NJ: Prentice Hall.

Endnote

[1]Random number generation was computed using the Text Mechanic™ Random Number Generator tool retrieved from http://textmechanic.com/text-tools/randomization-tools/random-number-generator/

Table 1
Probability terms in the lexicon and numeric probability equivalents

	Numeric Probability Equivalent		
Probability Term in Lexicon	Lexicon	Median	Frequency
Will not	0	0	61
No prospect	0	0.02	5
Little prospect	0.10	0.14	41
Extremely unlikely	0.10	0.07	37
Highly unlikely	0.10	0.10	37
Very unlikely	0.10	0.10	155
Low probability	0.25	0.25	80
Probably not	0.25	0.25	34
Unlikely	0.25	0.20	236
Slightly less than even chance	0.40	0.45	0
Even chance	0.50	0.50	5
Slightly greater than even chance	0.60	0.55	2
Probably	0.75	0.75	122
Probable	0.75	0.71	1
Likely	0.75	0.75	345
Highly likely	0.90	0.85	20
Extremely likely	0.90	0.90	0
Almost certain	0.90	0.95	41
Certain	1	1	17
Will	1	1	536

Note: Columns 1 and 2 show the probability lexicon in Barnes (2016). Column 3 shows the median numeric probability equivalent from Mandel (2015a). The last column shows the frequency of the terms in this research.

Table 2
Probability terms not in the lexicon and numeric probability equivalents

Probability Term	Reference Terms	NPE	Frequency
Very likely	Highly likely	0.85	64
Virtually certain	Almost certain	0.95	3
Inevitable	Almost certain	0.95	2
No chance	No prospect	0.02	4
Not likely	Unlikely	0.20	4
Is imminent	Certain	1	2
Inevitable	Certain	1	1
Chances are low	Low probability	0.25	1
Likelihood is low	Low probability	0.25	1
Not probable	Probably not	0.25	1
Very low probability	Very unlikely	0.10	1
Almost no prospect	No prospect, little prospect	0.05	14
Almost no chance	No prospect, little prospect	0.05	2
Very little prospect	No prospect, little prospect	0.05	2
Almost certainly not	(Almost certain)	0.03	2

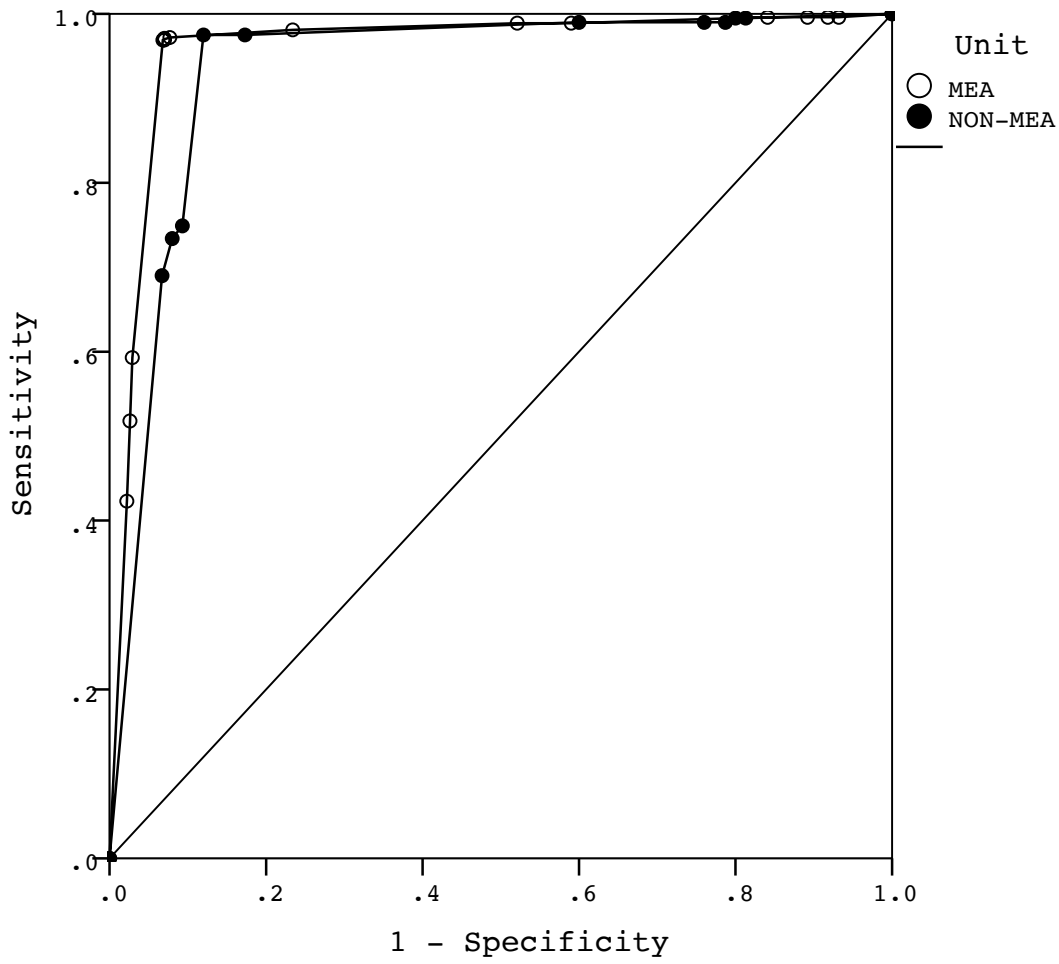

Figure 1. ROC curves by unit

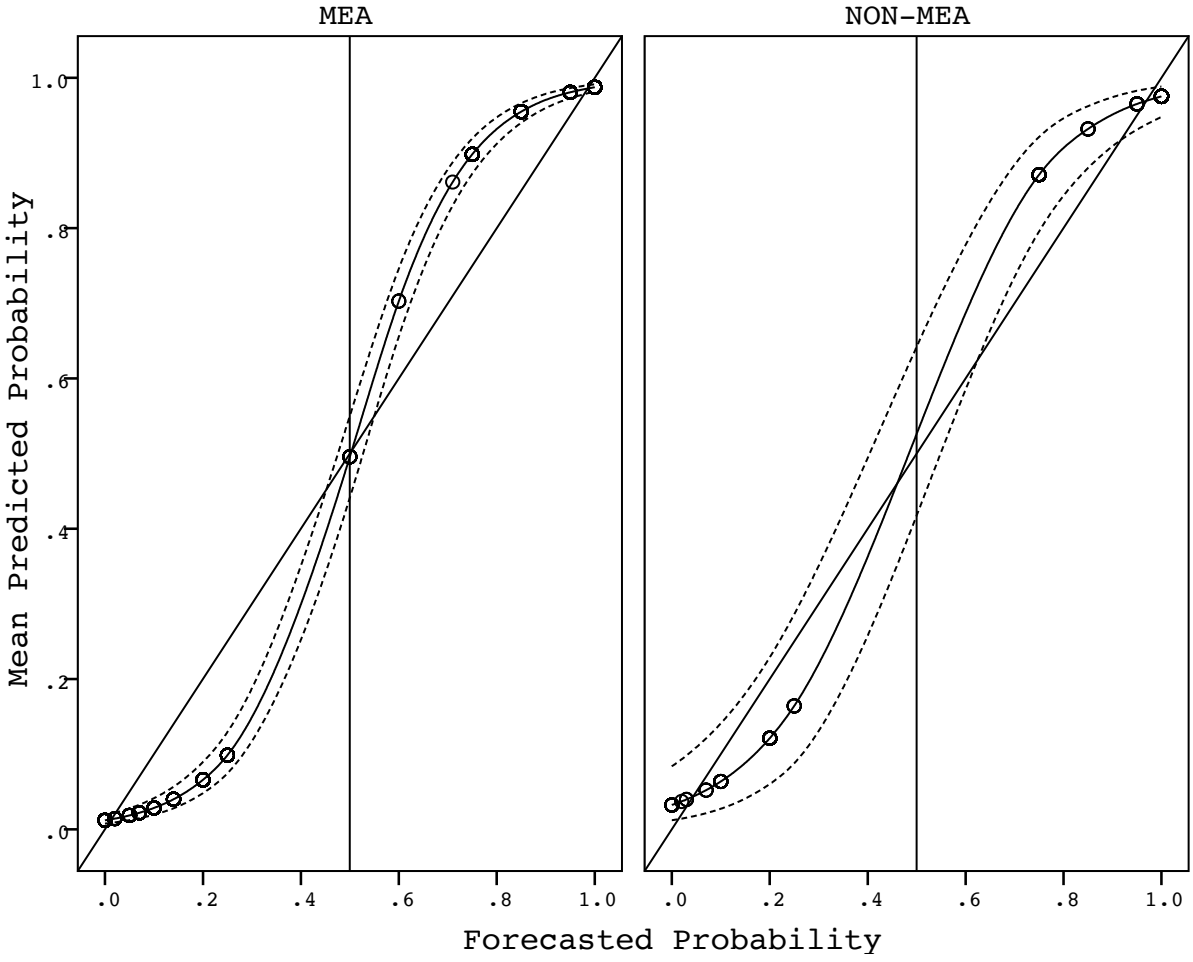

Figure 2. Model-based calibration curves by unit (dotted lines = model-based 95% CI)

DOCUMENT CONTROL DATA

(Security markings for the title, abstract and indexing annotation must be entered when the document is Classified or Designated)

1. ORIGINATOR (The name and address of the organization preparing the document. Organizations for whom the document was prepared, e.g., Centre sponsoring a contractor's report, or tasking agency, are entered in Section 8.)

DRDC – Toronto Research Centre
Defence Research and Development Canada
1133 Sheppard Avenue West
P.O. Box 2000
Toronto, Ontario M3M 3B9
Canada

2a. SECURITY MARKING (Overall security marking of the document including special supplemental markings if applicable.)

CAN UNCLASSIFIED

2b. CONTROLLED GOODS

NON-CONTROLLED GOODS
DMC A

3. TITLE (The complete document title as indicated on the title page. Its classification should be indicated by the appropriate abbreviation (S, C or U) in parentheses after the title.)

Geopolitical Forecasting Skill in Strategic Intelligence

4. AUTHORS (last name, followed by initials – ranks, titles, etc., not to be used)

Mandel, David R., Barnes, Alan

5. DATE OF PUBLICATION (Month and year of publication of document.)	6a. NO. OF PAGES (Total containing information, including Annexes, Appendices, etc.)	6b. NO. OF REFS (Total cited in document.)
November 2017	2	0

7. DESCRIPTIVE NOTES (The category of the document, e.g., technical report, technical note or memorandum. If appropriate, enter the type of report, e.g., interim, progress, summary, annual or final. Give the inclusive dates when a specific reporting period is covered.)

External Literature (P)

8. SPONSORING ACTIVITY (The name of the department project office or laboratory sponsoring the research and development – include address.)

DRDC – Toronto Research Centre
Defence Research and Development Canada
1133 Sheppard Avenue West
P.O. Box 2000
Toronto, Ontario M3M 3B9
Canada

9a. PROJECT OR GRANT NO. (If appropriate, the applicable research and development project or grant number under which the document was written. Please specify whether project or grant.)	9b. CONTRACT NO. (If appropriate, the applicable number under which the document was written.)

10a. ORIGINATOR'S DOCUMENT NUMBER (The official document number by which the document is identified by the originating activity. This number must be unique to this document.)	10b. OTHER DOCUMENT NO(s). (Any other numbers which may be assigned this document either by the originator or by the sponsor.)
DRDC-RDDC-2017-P091	

11a. FUTURE DISTRIBUTION (Any limitations on further dissemination of the document, other than those imposed by security classification.)

Public release

11b. FUTURE DISTRIBUTION OUTSIDE CANADA (Any limitations on further dissemination of the document, other than those imposed by security classification.)

12. ABSTRACT (A brief and factual summary of the document. It may also appear elsewhere in the body of the document itself. It is highly desirable that the abstract of classified documents be unclassified. Each paragraph of the abstract shall begin with an indication of the security classification of the information in the paragraph (unless the document itself is unclassified) represented as (S), (C), (R), or (U). It is not necessary to include here abstracts in both official languages unless the text is bilingual.)

Extending research by the authors on intelligence forecasting, the forecasting skill of 3,622 geopolitical forecasts extracted from strategic intelligence reports was examined. The codable subset of forecasts (N = 2,013) was expressed with verbal probabilities (e.g., likely) and translated to numeric probability equivalents. This subset showed very good calibration and discrimination, but also underconfidence. There was no support for the hypothesis that forecasting skill was good mainly due to the general ease of forecasting topics. First, forecasting skill was as good among authoritative key judgments as in the general set. Second, forecasts that were assigned high degrees of certainty, indicative of ease, ($P \leq 0.05$ or $P \geq 0.95$) did not discriminate as well as less certain forecasts ($0.05 < P < 0.95$), and these subsets did not differ in calibration. Sensitivity and benchmarking tests further revealed that if the 1,609 uncodable forecasts were all assigned forecast probabilities of .5 (i.e., if all followed a "cautious ignorance" rule), skill characteristics would still show a large effect size improvement over a variety of guesswork strategies. The findings support a cautiously optimistic assessment of forecasting skill in strategic intelligence and indicate that such skill is not primarily attributable to the selection of easy forecasting topics. However, the large proportion of uncodable cases suggests that intelligence forecasts could be improved by avoiding imprecise language that not only affects codability, but also, in all likelihood, the interpretability and indicative value of forecasts for intelligence consumers.

13. KEYWORDS, DESCRIPTORS or IDENTIFIERS (Technically meaningful terms or short phrases that characterize a document and could be helpful in cataloguing the document. They should be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location may also be included. If possible keywords should be selected from a published thesaurus, e.g., Thesaurus of Engineering and Scientific Terms (TEST) and that thesaurus identified. If it is not possible to select indexing terms which are Unclassified, the classification of each should be indicated as with the title.)

forecasting, prediction, intelligence analysis, skill, judgment