Forecasting the Accuracy of Forecasters from Properties of Forecasting Rationales

Christopher W. Karvetski[a,**] (ckarvetski@kadsci.com), Carolyn Meinel[b] (meinel@isitaustin.org), Daniel T. Maxwell[a] (dmaxwell@kadsci.com), Yunzi Lu[c] (yunzilu@sas.upenn.edu), Barbara A. Mellers[c] (mellers@wharton.upenn.edu), Philip E. Tetlock[c] (tetlock@wharton.upenn.edu)

[a] Kadsci, LLC, 4031 University Dr.Suite 100, Fairfax VA, 22030, USA

[b] 1234 Darless Drive, Cedar Park, TX 78613, USA

[c] Wharton School, University of Pennsylvania, Philadelphia PA 19104, USA

**corresponding author. E-mail address: ckarvetski@kadsci.com (Christopher W. Karvetski).

Abstract

Geopolitical forecasting tournaments have stimulated the development of methods for improving probability judgments of real-world events. But these innovations have focused on easier-to-quantify variables, like personnel selection, training, teaming, and crowd aggregation—and bypassed messier constructs, like qualitative properties of forecasters' rationales. Here we adapt methods from natural language processing (NLP) and computational text analysis to identify distinctive reasoning strategies in the rationales of top forecasters, including: (a) cognitive styles, such as dialectical complexity, that gauge tolerance of clashing perspectives and efforts to blend them into coherent conclusions; (b) the use of comparison classes or base rates to inform forecasts; (c) metrics derived from the Linguistic Inquiry and Word Count (LIWC) program. Applying these tools to multiple forecasting tournaments and to forecasters of widely varying skill (from Mechanical Turkers to carefully culled "superforecasters") revealed that: (a) top forecasters show higher dialectical complexity in their rationales, use more comparison classes, and offer more past-focused rationales; (b) experimental interventions, like training and teaming, that boost accuracy also influence NLP profiles of rationales, nudging them in a "superforecaster-like" direction.

Keywords: geopolitical forecasting, psycholinguistic vocabularies, integrative complexity, comparison class, natural language processing, LIWC

1. Introduction

Researchers have used forecasting tournaments to answer questions posed by a wide range of professional-scientific communities, including psychologists specializing in judgment under uncertainty (Atanasov et al., 2020; Baron et al., 2014; Bo et al., 2017; Chen et al., 2016; Mellers et al., 2015; Moore et al., 2017); political scientists and economists curious about limits of the predictability of complex social systems (Baron et al., 2014; Friedman et al., 2018; Lustick & Tetlock, 2020; Scoblic & Tetlock, 2020; Tetlock, 2017); statisticians developing algorithms for distilling wisdom from crowds (Cross et al., 2018; Satopää et al., 2014); and intelligence analysts charged with anticipating threats to national security (Chang et al., 2016, 2017). This burgeoning literature has led to the discovery of methods for improving probability judgments of real-world events often thought "too unique" and thus beyond probability estimation, including psychometric tests for screening talent (Mellers et al., 2015), debiasing training (Chang et al., 2016), team exercises for facilitating constructive debate (Tetlock & Gardner, 2015), and aggregation algorithms (Baron et al., 2014; Cross et al., 2018; Satopää et al., 2014). Researchers have also made progress modeling the statistical pathways showing how various methods enhance forecasting performance: improving signal extraction, reducing systematic biases, and tamping down noise (Satopää et al., 2020).

This article fills a conspicuous gap in this research literature: the lack of attention to the rationales that forecasters construct in support of their forecasts. The omission is understandable, though not justifiable. Free-flowing natural-language data are messy: sentence fragments, dangling references, idiosyncratic jargon, tangential digressions, and so on. Nonetheless, it would be surprising if there were no systematic linkages between how forecasters explain their forecasts and the accuracy of those forecasts, especially in datasets large enough to tamp down

noise. Consistently null results would suggest that we inhabit a roulette-wheel universe that forecasters stubbornly refuse to accept—and persist in churning out specious explanations for illusory patterns (Tetlock, 2017, Chapter 3).

We do not see the search for systematic links as a fool's errand. Drawing on the Kahneman (2011) heuristics-and-biases tradition as well as a tradition of work on cognitive-style correlates of good judgment (Suedfeld & Tetlock, 2001), we conjecture that geopolitical forecasters will be more successful to the degree they resist two potent temptations: (a) basing their probability judgments exclusively on an inside view of the case at hand and failing to step back and take the outside view (locating the case in comparison classes that permit estimates of base rates of how often things of this sort happen in situations of this sort); (b) constructing rationales for their probability estimates in ways that minimize the strain of cognitive dissonance and yield neat narratives about the inevitability of outcomes while failing to recognize that opposing perspectives often contain key kernels of truth.

Turning first to the comparison-class hypothesis, consider a reasonably representative question posed in a 2019 geopolitical forecasting tournament sponsored by the U.S. intelligence community, Intelligence Advanced Research Projects Activity (IARPA):

"Will North Korea launch a medium-range or longer ballistic missile between 20 June 2019 and 30 September 2019?"

As is common in such tournaments, forecasters offered a wide range of rationales for their forecasts. One took the inside view and tried to peer into the North Korean dictator's mind: "When Kim feels ignored he goes to great lengths to get the attention of everyone in the room." A second took the outside view and assimilated the case at hand to a comparison class: "The last known case of a medium or longer missile launch from North Korea was in 2017." The first

forecaster speculated about Kim's state of mind; the second observed an event frequency that could become the basis for a rough probability estimate.

Of course, there is no guarantee that the outside view will always yield superior accuracy. Shrewd insights into mindsets of outlier decision makers will sometimes trump crude comparison classes that lump together cases that do not belong together. But a sizable body of work on judgment under uncertainty (Kahneman, 2011) as well as on forecasting tournaments in particular (Tetlock, 2017; Tetlock & Gardner, 2015) suggests that betting on the outside-view forecasters is likelier to pay off in the long run—and the second forecaster will prove more accurate not just on this question but on a variety of others. The mental habit of adopting the outside view—constructing comparison classes and estimating base rates—is a foundational forecasting skill (Armstrong, 2001, 2005)—and indeed served as the first of the ten training checklist items known by the acronym "CHAMPS KNOW," where "C" stood for comparison class (Chang et al., 2016; Tetlock & Gardner, 2015; for a full description, see Appendix A). CHAMPS KNOW was developed and refined in the first wave of IARPA tournaments in 2010-15. Chang et al. (2016) demonstrated that CHAMPS KNOW training improved forecasting accuracy. Furthermore, the more frequently forecasters reported using the "C" component of the training, the better they did.

The Chang et al. (2016) evidence is far from decisive, however. They never directly measured whether forecasters actually used comparison classes when making their probability judgments. They relied on box-checking, self-reports after the fact. Retrospective self-reports of judgment strategies are highly imperfect gauges of what people were once really thinking (Nisbett & Wilson, 1977) or even of what they may have said in their original rationales. To overcome these limitations, we developed a text-based classification model to automate detection

of comparison classes in forecasting rationales. The model allows us to test our first set of hypotheses about individual-differences and situational variations in forecasting skill. Hypothesis 1 consists of two parts: (a) Forecasters with consistently better accuracy track records (e.g., "superforecasters") will invoke comparison classes more often in their rationales than do less accurate forecasters; (b) forecasters working under experimental conditions known to boost accuracy (training and teaming interventions) will invoke comparison classes more often than forecasters not assigned to these treatment conditions.

Our second set of hypotheses bears on how forecasters integrate evidence and arguments to generate forecasts as well as their capacity to cope with complexity and dissonance (Jervis, 1997). Here we borrow and adapt assessment tools from the integrative-complexity (IC) research program that dates back to Schroder, Driver, and Streufert (1967) and that encompasses a variety of laboratory and archival studies of the correlates of judgmental accuracy (Suedfeld & Tetlock, 1977, 2001; Tetlock & Kim, 1987). IC has been operationalized on a seven-point scale for assessing awareness of alternative perspectives and for connecting perspectives to reach integrative conclusions (for more detail, see Appendix B; for overviews of applications, see Conway et al., 2014; Suedfeld & Tetlock, 2014). Given the repeated demonstrations that manipulations that increase integrative complexity tend to decrease over-confidence (Tetlock & Kim, 1987), we hypothesize that more accurate forecasters will have higher levels of IC, and that also that teaming of forecasters (as opposed to working independently) will enhance forecaster accountability (Lerner & Tetlock, 1999), expose forecasters to more viewpoints (Page, 2018) and increase the level of IC within rationales.

The work of Conway et al. (2008) and of Tetlock and Tyler (1996) does however suggest an important refinement of this argument. These authors distinguish two types of complexity:

dialectical complexity, which involves grappling with the cognitive tensions between competing perspectives (more "howevers"), and *elaborative* or cognitive complexity, which involves reducing tensions by generating reinforcing reasons for taking strong stands (more "moreovers"). Conway et al. (2020) found that dialectical complexity was negatively correlated with confidence and extremity in one's point of view whereas elaborative complexity was positively correlated. These findings suggest that higher dialectical but not elaborative complexity will predict better forecasting—because over-confidence is more common than under-confidence in geopolitical forecasting (Tetlock, 2017; Tetlock & Gardner, 2015). Hypothesis 2 also consists of two parts: (a) Forecasters with better accuracy track records (e.g., "superforecasters") will generate more dialectically complex rationales for their forecasts than will less accurate forecasters; (b) forecasters working under experimental conditions known to boost accuracy (training and teaming interventions) will generate more dialectically complex rationales than forecasters not assigned to these treatment conditions.

Although our central hypotheses focus on comparison classes and IC, we also build on Zong et al. (2020) who studied forecaster rationales drawn from Good Judgment Open[1], an open source forecasting platform. Using Linguistic Inquiry and Word Count, a widely used psycholinguistic system (Pennebaker et al., 2015; Tausczik & Pennebaker, 2010), as well as other NLP variables such as readability, Zong et al. (2020) found that better forecasters: (a) scored higher on "tentativeness"[2]; (b) emphasized the past more than the present or future; (c) quoted more from outside sources, a sign of more aggressive information search; (d) used more first-person than third-person pronouns, perhaps a sign of willingness to take responsibility for conclusions (Kacewicz et al., 2014). Our study provides a chance to cross-validate the Zong et

[1] https://www.gjopen.com/
[2] Examples from this category include words like "seems", "maybe", "perhaps".

al. (2020) findings on a larger forecaster population and larger set of questions. Instead of relying on a single tournament, we explore patterns across multiple IARPA tournaments, spanning a decade, and on forecasters ranging in skills from Mechanical Turkers to "superforecasters." The tournaments used here also randomly assigned participants to experimental conditions, including training, teaming, and process accountability (a condition in which forecasters expected to be evaluated not on accuracy but on how well their rationales adhered to CHAMPS KNOW training guidelines, Chang et al., 2017). Our database thus offers opportunities for testing a wider range of hypotheses.

In sum, the current research advances knowledge in three ways: first, by using natural language methods to test hypotheses about comparison classes and accuracy (Chang et al., 2016); second, by exploring links between integrative/dialectical complexity and accuracy; third, by connecting to the LIWC and in doing so assessing the replicability of Zong et al. (2020).

2. Method and Metrics

2.1. Overview of forecasting tournaments

IARPA's geopolitical forecasting tournaments between 2010 and 2020 have posed hundreds of well-defined forecasting questions to thousands of forecasters. The first wave of tournaments, known as Aggregative Contingent Estimation (ACE), stretched over four years, from 2010 to 2015, with intermittent tournaments thereafter. Questions in each tournament lasted from several weeks to upwards of a year. All questions featured binned choices, with only one bin eventually resolving as "true". Some questions had binary "yes"/"no" bins: "Will the United Kingdom (UK) leave the European Union (EU) before 1 November 2019?" Some had multinomial bins where bin order was not meaningful (e.g., an election with more than two candidates). Still others had

multinomial bins where bin order mattered (e.g., "What will be the daily closing price of gold on 30 October 2019 in USD?", with bins representing discretized intervals for prices).

Table 1 illustrates the broad spectrum of topics covered in a recent IARPA tournament, the Geopolitical Forecasting Challenge 2 (GFC2), open between May and November, 2019.[3] Each forecaster would provide a probability distribution over the answer bins and a text-based rationale. Forecasters were encouraged to update forecasts and rationales as new information became available or as deadlines for resolution approached. Forecasts for unordered questions, binary or multinomial, were evaluated using Brier scoring (Brier, 1950), a *proper scoring rule* that incentivizes truthful reporting (Gneiting & Raftery, 2007) and considers both discrmimination and calibration (Yaniv et al., 1991). Ordered questions were evaluated by a squared scoring rule sensitive to distance (Jose et al., 2009).

To control for variation in question difficulty, we standardized forecasters' scores. Standardization takes into account that a close-to-zero Brier score on questions where everyone else is close to zero is far less impressive than close-to-zero score on questions where everyone else is struggling and getting poor scores (indicating a big gap between probability judgments and reality). For more detail on standardization, see the data description section for each analysis.

Depending on the assignment of experimental conditions, forecasters worked independently, collaboratively in teams, or in a mixed environment where they could see rationales and forecasts of others but not interact. Depending on condition, rationales accompanying each forecast included the forecaster's line of reasoning, posting of news stories and links, or intra-team messaging.

[3] Three of the authors competed within the tournament as members of Team KaDSci, placing 4th overall out of 36 teams.

Table 1. Summaries of the diverse topics covered in a recent IARPA forecasting tournament (GFC2) which took place in 2019.

Domain	N (%)	Examples
Health/ Disease	31 (8%)	• Will there be more than 3,500 cumulative Ebola cases from the North Kivu Ebola Outbreak before 1 October 2019, according to the Humanitarian Data Exchange? • Before 1 July 2019, will the FAO report 1,200,000 or more pigs have been culled in China to thwart the spread of African Swine Fever?
Macroeconomics/ Finance	55 (14%)	• Will there be a 10% decrease in the daily closing price of the FTSE 100 over any five (5)-business day interval between 2 May 2019 and 20 November 2019? • According to the Council on Foreign Relations (CFR) Global Monetary Policy Tracker, what will be the global monetary policy index in October 2019?
Natural sciences/ Climate	19 (5%)	• Will there be an earthquake of magnitude 8.5 or stronger worldwide between 23 May 2019 and 31 October 2019? • According to the Global Disaster Alert and Coordination System (GDACS), will there be any Red Alerts for droughts for Thailand between 18 July 2019 and 29 November 2019?
Politics/ International Relations	279 (70%)	• Between 16 May 2019 and 30 July 2019, will the president of the Democratic Republic of the Congo (DRC) appoint a prime minister? • Will Hong Kong's Chief Executive Carrie Lam experience a significant leadership disruption between 27 June 2019 and 29 November 2019?
Technology	15 (4%)	• Will a nuclear or radiological event occur with an INES Level 3 or higher rating between 30 May 2019 and 31 August 2019? • Before 30 July 2019, will a merger, acquisition, and/or joint venture agreement be announced between Telecom Italia and Open Fiber?

2.2. Overview of methods and metrics

Forecasters often generated multiple rationales per question so we decided to select the forecaster's first rationale, when the forecaster is initially confronting the problem rather than updating a prior thought. Within a tournament, we then selected the subset of forecasters who

made predictions for at least ten questions and offered rationales of at least ten words—to ensure a substantial corpus of rationales for each forecaster and to stabilize our standardized accuracy metric. We scored the selected rationales for IC using the Conway et al. (2014) and Houck et al. (2014) autoIC tool, which yields an *Integrative Complexity* score as well as scores for *Dialectical Complexity* and *Elaborative Complexity*. Each of the three scores range from 1 to 7.

We also evaluated rationales on use of comparison classes. Unlike IC, where an existing tool was available, we had to develop and validate our own method. We created a paired sample by selecting two rationales from each of 100 selected questions from the GFC2 forecasting tournament, with one rationale featuring and the other not featuring a comparison class. We then trained a random forest (RF) model on the terms most predictive of rationales using a comparison class. Figure 1 shows a word-cloud from this RF model, with the more prominent words receiving the highest "importance score" within the RF training. Prominent words include references to time (e.g., "year", "month", "last", "past"), as well as to data and statistics (e.g., "data", "average", "ranged", "highest") and relativity (e.g., "between", "below", and "than"). Using this model, we assigned each rationale a *Comparison Class* score between zero and one reflecting the model-assigned probability that the rationale features a comparison class. Full validation details are in Appendix C.

We then generated a LIWC profile using the 2015 version of the software and selected variables (from over 90 in the output file) to examine correlations of these variables with overall performance, as well as comparison-class usage and integrative complexity scores. Each LIWC variable consists of a numeric score between zero and 100, with larger values indicating more terms linked to the variable topic appearing in the rationale. LIWC variables examined include tentativeness, use of first-person singular, third-person singular, and third-person plural

pronouns, focus on the past, focus on the present, focus on the future, and quotes. We also examined informality to measure engagement among teammates. The first column of Table 2 lists the categories; the second column lists the variable names that match with Pennebaker et al. (2015); the third column lists examples.

Figure 1. Word cloud of terms most predictive of rationales featuring comparison classes.

Table 2. Summary of LIWC categories used within our analyses.

Category	LIWC Variable	Examples
Tentativeness	*tentat*	"maybe", "perhaps", "seems"
First Person Singular	*i*	"I", "me", "my"
Third Person Singular	*shehe*	"she","her","him"
Third Person Plural	*they*	"they", "their"
Focus on Past	*focuspast*	"did", "ago"
Focus on Present	*focuspresent*	"is", "now"
Focus on Future	*focusfuture*	"will", "soon"
Use of Quotes	*quotes*	(presence of quotes)
Informal Words	*informal*	"agree", "haha", (emoticons)

We counted web links posted within each rationale, labeled *Source Count,* to measure the extent to which forecasters used outside sources. We also included *Word Count*, which has emerged as a weak but consistent correlate with accuracy in past work. We tested our hypotheses by averaging linguistic-rationale metrics across questions for each forecaster and correlating averages with forecasters' standardized accuracy scores to determine which variables distinguished better from worse forecasters. Study 2 had experimental assignments to conditions, so we could also investigate effects of the interventions on standardized performance and rationale metrics.

3. Study 1

3.1. Data set

Our first analysis used data from the GFC2 forecasting tournament in which IARPA posted 399 questions at weekly intervals and provided a daily stream of forecasts from 537 individuals

contracted through the Mechanical Turk platform[4]. These "Turker" forecasters (henceforth described as "forecasters") were not working in teams but could view the rationales and forecasts of their peer forecasters. In total, they provided 75,479 forecasts. The average time a question was open was M = 97.2 days.

IARPA created a benchmark based on a consensus model from forecasters' stream of forecasts. The model produced an updated and aggregated numeric forecast for each question for each day the question was "live" and incorporated extremizing and other techniques to maximize accuracy (see Baron et al., 2014 for a similar model). We compared forecasts to the same-day consensus forecasts as a measure of standardized accuracy. We scored all forecasts and corresponding consensus forecasts using Brier scores (ordinal or non-ordinal) with an R package (Merkle & Steyvers, 2013)[5], where scores were set to range from 0 (best) to 1 (worst). We call the score of a forecaster *Forecast Score* and the score of the corresponding consensus forecast (same day/question forecast from consensus model) *Consensus Score*. We eliminated forecasts and rationales when the corresponding Consensus Score was less than or equal to 0.0025[6] in order to filter out forecasts submitted near the end of a question's designated timeline, when resolution was a virtual certainty, leaving 44,748 forecasts (out of 75,479). We also eliminated rationales that contained fewer than ten words and restricted our analyses to forecasters with ten-plus forecasts and rationales, yielding a final sample 27,507 forecasts and rationales from 485 forecasters (an average of about 57 questions per forecaster).

[4] Teams were not required to use the Turkers' forecasts in their solutions. All data (question metadata, forecasts, rationales, etc.) were provided by IARPA using APIs, with each forecaster identified by a random number. Forecaster data were anonymized before distribution to the teams with no indication to the identities of forecasters.
[5] https://cran.r-project.org/web/packages/scoring/scoring.pdf
[6] This is equivalent to the score one would get if one put a 0.95 probability on a binary event that resolved as true.

To measure the relative performance of forecasters, we calculated the difference between the Consensus Score and Forecast Score. To get an overall standardized measure of accuracy for each forecaster$_i$ who made $J_i \geq 10$ total forecasts, we averaged the score differences as follows:

$$\Delta_i = \sum_{j=1}^{J_i} \left(Consensus\ Score_j - Forecast\ Score_j \right) / J_i = \sum_{j=1}^{J_i} \Delta_{i,j} / J_i, \quad (1)$$

where Δ_i represents the average score difference for forecaster$_i$. Δ_i was positive when the Consensus Score was larger (or worse) than the Forecast Score, implying the forecast was more accurate than the consensus. The mean Δ_i across all 485 forecasters in the sample was M = -.01 with a 95% confidence interval of [-.01, -.01]. Figure 2 shows the histogram of averaged differences for forecasters (Δ_i's), with a distribution centered near zero, a sign that the consensus model is a suitable benchmark. We scored these 27,507 rationales on IC, comparison class, and the LIWC variables. The average number of words per rationale was M = 73.4, [73.0, 73.9].

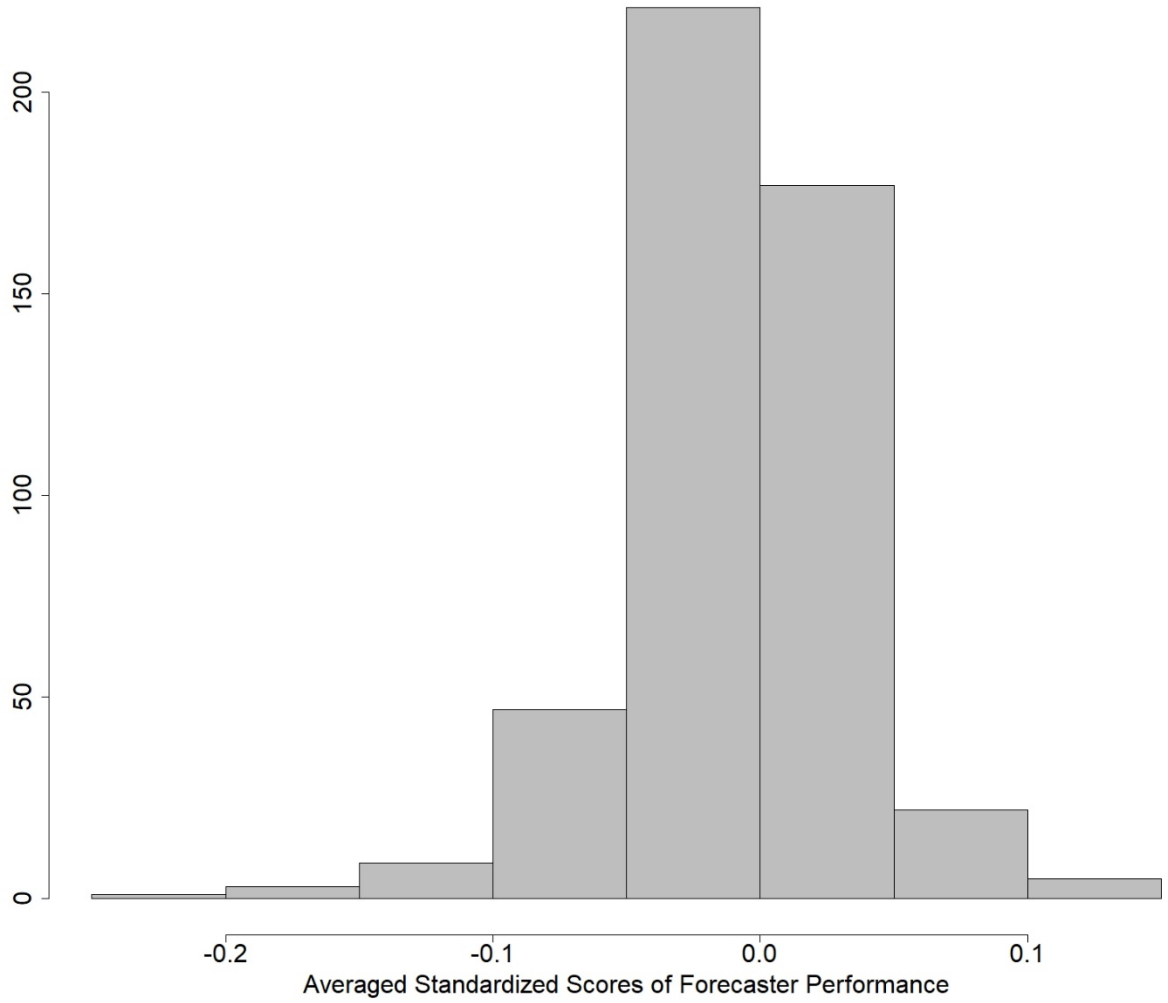

Figure 2. Histogram of Average Differences between Consensus and Forecaster Scores (Δ_i's) for 485 forecasters.

3.2. Study 1 results

To collect information for each forecaster, we averaged the rationale-based metrics (e.g., Integrative Complexity, Comparison Class, LIWC variables) for each individual. To test our hypotheses that better forecasters were likelier to consider comparison classes and have higher IC, and to test the replicability of other researchers' findings (e.g., better forecasters are more past-focused), we correlated each average variable with Δ_i accuracy values. Figure 3 shows these

correlations, with the key result in the first column where Δ_i is the *Standardized Accuracy Score*. As expected, Δ_i correlates weakly with Word Count, $r_{wc} = r(485) = .12$, $p = .007$, and Word Count is highly correlated ($r > .70$) with Comparison Class and the three IC variables.

The only variable significantly negatively correlated with Δ_i was Focus on the Future, $r(485) = -.09$, $p = .04$. Table 3 shows variables significantly positively correlated with Δ_i. Using the method[7] outlined by Steiger (1980) and Lee and Preacher (2013), we examined which of these correlations were greater than r_{wc}. Results appear in the last column of Table 3. Comparison Class, Integrative Complexity, and Dialectical Complexity were the only variables with significantly greater correlations with Δ_i than Word Count.

Overall, the findings support both Hypotheses 1 and 2: better forecasters were likelier to use comparison classes, and express higher Integrative Complexity and Dialectical Complexity, both of which were better predictors than Elaborative Complexity. We find mixed support for the hypotheses of Zong et al. (2020). Tentativeness and Focus on the Past were associated with better performance, but Use of Quotes and First-person Singular Pronouns were not.

Table 3. Positive correlates with overall forecasting accuray (Δ_i).

Indicator	Correlation with Δ_i	p-value; $r \neq 0$	p-value; $r > r_{wc}$
Word Count	.12	**	--
Comparison Class	.18	***	*
IC	.17	***	*
DIAL	.20	***	**
ELAB	.11	*	--
Tentativeness	.17	***	--
Focus on the Past	.13	**	--

^ $p < .1$, * $p < .05$, **$p < .01$, *** $p < .001$.

[7] The method tests the equality of two correlation coefficients obtained from the same sample, with the two correlations sharing one variable in common.

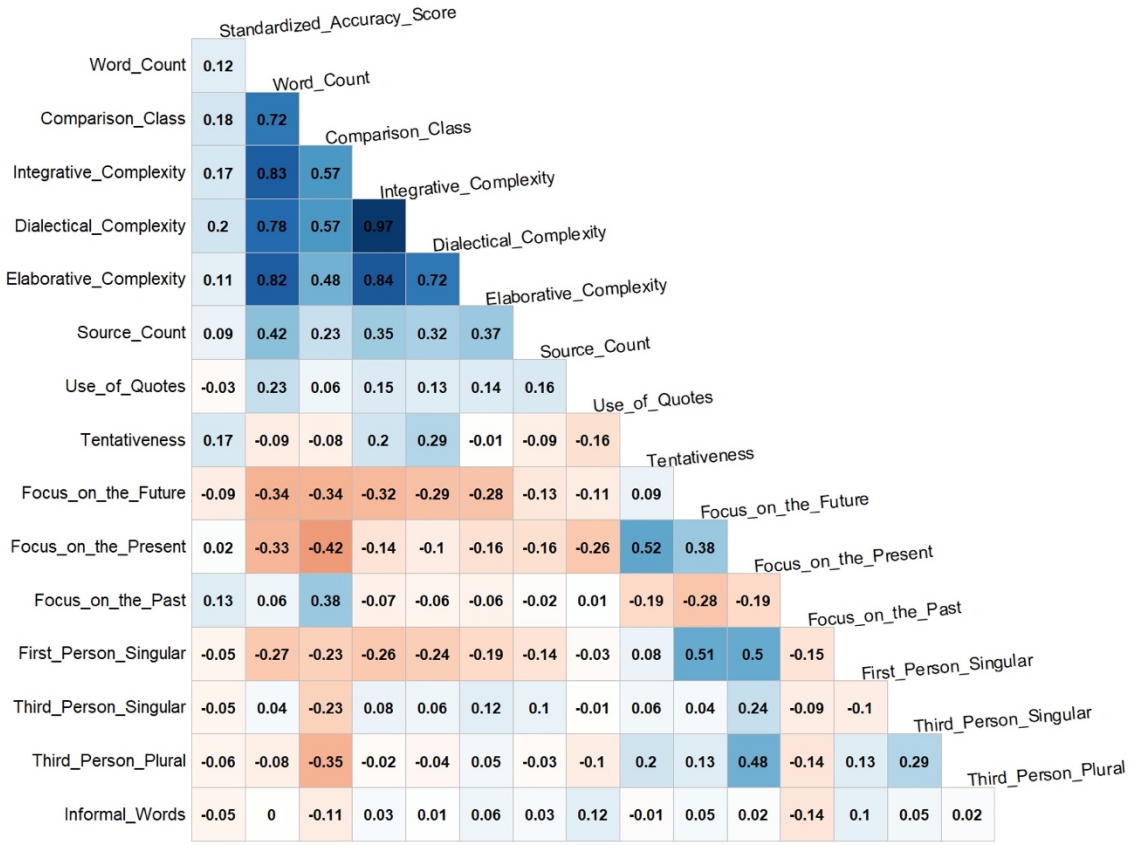

Figure 3. Correlations among Δ_i values (called Standardized Accurary Score), Comparison Class, Integrative Complexity, and LIWC variables. Darker blue indicates a greater positive association; darker red, a greater negative association.

4. Analysis Study 2

4.1. Data set

Our second analysis drew on data from the Good Judgment Project in the ACE tournament, which ran from 2010 to 2015 and has inspired many publications (Chang et al., 2016, 2017; Chen et al., 2016; Horowitz et al., 2019; Mellers et al., 2014, 2015, 2017; Satopää et al., 2014; Schwartz et al., 2017; Tetlock et al., 2014; Tetlock & Gardner, 2015). In year 1, Good Judgment randomly assigned 2400 forecasters to one of 12 conditions in a 3 x 4 factorial design, with roughly 200 subjects per condition: Training Conditions (No Training, Probability Training, and

Scenario Training) x Elicitation Conditions of Control (independent forecasters), Crowd Beliefs (independent forecasters who saw the distribution of others' forecasts but could not communicate), Teams (forecasters who worked in groups of 15-20 and were asked to justify their forecasts to each other) and a prediction market (in which rationales played less of a role and will not be discussed further).

In year 2, 1860 forecasters were randomly assigned to a 2 x 3 Training-by-Elicitation factorial design, with roughly 300 forecasters per condition, plus an offset special condition of 60 superforecasters who were top performers in Year 1. Levels of training were "No Training" and "Probability Training", which included discussions about constructing reference classes, averaging multiple predictions from different sources, and avoiding judgmental biases such as over-confidence, confirmation bias, or base-rate neglect. Levels of Elicitation were a Control Group of forecasters working independently, Teams who worked in groups of roughly 15 forecasters, and Continuous Double Auction Prediction Markets (not discussed further).

Year 3 had approximately 3,000 forecasters—plus an offset condition of 120 superforecasters based on past performance. Good Judgment randomly assigned 600 forecasters to a condition in which they worked alone, equally divided into probability training and no-training sub-conditions. Good Judgment also assigned 750 forecasters to teams (375 with team facilitators and 375 without, a distinction we do not discuss here). And Good Judgment assigned the remainder of forecasters to three prediction markets (an LMSR market, a CDA market with 550 individuals, and a CDA market, not mentioned again here).

In Year 4, Good Judgment assigned 12,280 forecasters to the following experimental conditions: independent, separated into training and no training sub-conditions; teams who worked with or without facilitators sub-conditions and prediction markets (not discussed). Year 4

also had a separate side-experiment in which roughly 2000 forecasters were assigned to one of three conditions: process accountability (forecasters were judged by how well their rationales adhered to CHAMPS KNOW training guidelines); outcome accountability (forecasters were judged on their accuracy or Brier scores); and a hybrid process-outcome accountability (forecasters were judged by both process and outcome standards). See Chang et al. (2017) for more details. Finally, Year 4 also had an offset condition of "superforecasters" culled from the previous three years, now numbering approximately 180. Given the top performers worked together in teams of 12 each year, their superior accuracy was likely driven by a mix of personnel selection and team dynamics.

Each year, forecasters could answer as many questions as they wished. We standardized each forecaster's score on a question by subtracting the score from the mean and dividing by the standard deviation, in keeping with the convention that higher standardized scores imply better performance. These standardized scores were then averaged over questions within a year to yield a *Standardized Yearly Score*. We restricted our analysis to forecasters making at least ten predictions in a year and offering rationales of ten-plus words.

Table 4 shows the number of forecasters who met these criteria, by treatment condition and year, with averaged Standardized Yearly Score as well as Word Count. Overall, there were 1,948 forecaster-year combinations, with 69,263 (initial) forecasts across 481 unique questions. In year 4, forecasters also rated their rationales based on whether they considered a comparison class using a check-box, which provided an instructive contrast with automated model-assessed Comparison Class variable which was developed using data from the GFC2 (Study 1).

Table 4. Statistics for Study 2.

Year	Condition	# of Forecasters	Avg. # of Questions	Avg. Std Yearly Score	Avg. Word Count
1	Individuals no training	14	14.3	-0.17	28.6
1	Individuals training	9	14.4	0.02	26.2
1	Team no training	20	13.7	0.14	41.9
1	Team training	26	14.7	0.24	45.5
2	Individuals no training	53	28.7	-0.17	24.0
2	Individuals training	39	37.2	0.08	27.7
2	Team no training	87	27.4	0.10	41.3
2	Team training	102	31.3	0.17	39.2
2	Supers Team training	38	65.4	0.45	60.5
3	Individuals no training	41	28.6	-0.17	36.0
3	Individuals training	45	22.1	-0.13	46.8
3	Team training	326	33.6	0.13	52.5
3	Supers Team training	100	54.8	0.38	64.7
4	Individuals no training	107	30.1	-0.07	28.4
4	Individuals training	92	30.8	0.01	45.9
4	Individuals training, hybrid accountability	82	29.2	0.09	57.5
4	Individuals training, process accountability	82	28.4	0.00	61.9
4	Team training	208	38.6	0.21	53.6
4	Team training, hybrid accountability	174	35.5	0.22	68.4
4	Team training, process accountability	187	34.1	0.18	73.2
4	Supers Team training	116	62.9	0.39	83.2

4.2. Results

4.2.1. Across treatment conditions

We began by testing the effects of training, teaming, superforecasting and process accountability on performance, as done in other published accounts, such as Mellers et al. (2014). We looked at effects of the interventions on Word Count and Standardized Yearly Score, both of which were the dependent variable in a regression model with interventions denoted as dummy variables. Table 5 presents the results. Using untrained individuals working independently as the baseline, process accountability had only a weak effect on accuracy. Training provided roughly twice the benefit of process accountability. Teaming and superforecasting conferred even larger benefits. Again, using untrained individuals working

independently as the baseline, teaming added about 9 words, training, about 16 words, process accountability, about 21 words, and superforecasting, about 26 words.

Table 5. Regression coefficients of Interventions on Standardized Yearly Scores and Word Count.

	Dependent Variable	
	Standardized Yr. Score (N = 1,948)	Word Count (N = 69,263)
Intercept	-0.11***	31.9***
Training	0.09***	15.9***
Teaming	0.18***	8.5***
Superforecasting	0.23***	25.7***
Process Accountability	0.04**	20.6***

^ p < .1, * p < .05, **p < .01, *** p < .001.

To examine the correlations of rationales with performance, we averaged the rationale-based linguistic variables for each forecaster in a given year and correlated them with each forecaster's Standardized Yearly Score. Figure 4 shows correlations over four years, with the first column being Standardized Yearly Scores. All variables, except Tentativeness, ($r(1,948) = .03$, $p = .17$) were significantly correlated with Standardized Yearly Score. In particular, Comparison Class had the largest correlation with Standardized Yearly Score, ($r(1,948) = .32$), followed by Integrative Complexity and Dialectical Complexity, each being $r(1,948) = .28$. Only these three variables exhibited significantly greater correlations with Standardized Yearly Score than Word Count ($r > r_{wc}$) at $\alpha = .05$ or better, with $p < .0001$. Consistent with Study 1, the results support both Hypotheses 1 and 2 that better forecasters are likelier to use comparison classes and have higher dialectical complexity. Aside from the finding on Tentativeness, the LIWC variables closely align with the results of Zong et al. (2020). Better forecasters used more sources and quotes, were less focused on the future, less focused on the present, more focused on the past, and used more first-person than third-person pronouns.

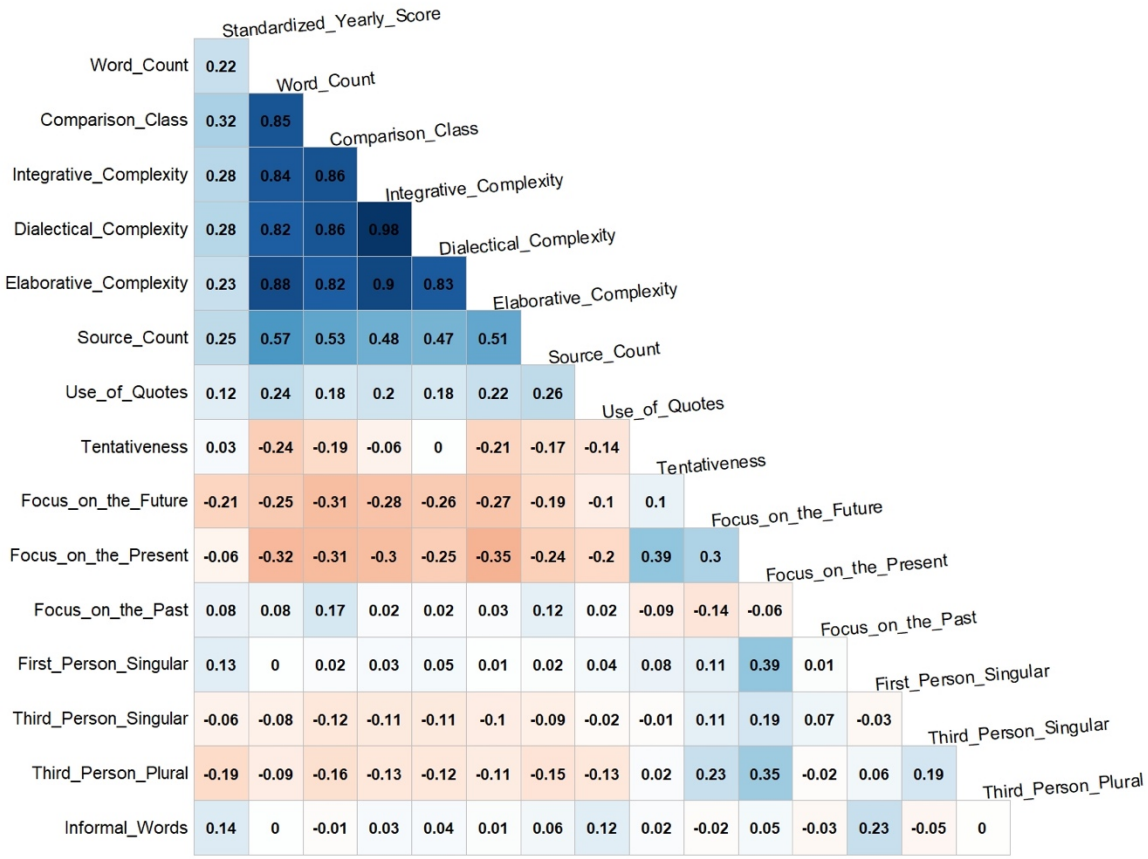

Figure 4. Correlations of Standardized Yearly Scores, Comparison Class, Integrative Complexity, and LIWC variables.

We compared self-assessed use of comparison classes to NLP-assessed use of comparison classes in year 4. The correlation of NLP-assessed Comparison Class with Standardized Yearly Score was r(1,048) = .33, and the correlation between the average self-assessed values and Standardized Yearly Score was r(1,048) = .15. The two comparison class variables, self-assessed and model assessed, were correlated by only r(1,048) = .29.

To explore relationships among treatment conditions and linguistic-rationale variables, we used Least Absolute Shrinkage and Selection Operator (LASSO) regressions with each rationale-based variable (i.e., Comparison Class, IC and LIWC variables, etc.) as the dependent variable and treatment interventions as dummy predictor variables. Rather than using averaged values for

each linguistic variable, we used the entire set 69,263 individual rationales to focus on rationale-level differences while controlling for Word Count, which is included as an explanatory variable. We used LASSO regressions to exclude coefficients with minimal effects (i.e., "variable selection". See Hastie et al., 2017 for a general overview of LASSO regression)[8].

Table 6 shows that Comparison Class was predicted by each of the four interventions with similar marginal contributions. Integrative Complexity and Dialectical Complexity were most associated with the superforecasting condition, followed closely by teaming, then training, and process accountability. The ordering of interventions was similar for Elaborative Complexity, but the difference between superforecasting and process accountability was smaller. With Source Count, superforecasting and teaming interventions had the largest marginal contributions. Superforecasters were likelier to provide quotes in their rationales. Finally, with Tentativeness, all interventions but teaming had negative marginal contributions, with training having the largest negative marginal effect.

Whereas training, teaming and superforecasting were associated with a decrease in the tendency to focus on the future, process accountability was associated with a slight increase. Interventions had mixed effects on Focus on the Present and Focus on the Past. With Focus on the Past, superforecasting was the only intervention associated with a positive effect. Both the teaming and superforecasting interventions were correlated with using First-Person Singular Pronouns, whereas all four interventions were correlated with using less Third-Person pronouns. Finally, teaming and superforecasting were correlated with use of Informal Words; process accountability was correlated with a decrease in Informal Words.

[8] We use glmnet package in R to perform all LASSO regressions ($\alpha = 1$) and set lambda to lambda.min from the cv.glmnet routine in order to select out final coefficients.

Table 6. LASSO regression coefficients for the interventions on each linguistic variable.

		Interventions				
Dependent variable	Intercept	Word Count	Training	Teaming	Supers	Process Account
Comparison Class	0.17	0.001	0.02	0.02	0.02	0.02
Integrative Complexity	1.42	0.007	0.06	0.09	0.11	0.03
Dialectical Complexity	1.34	0.006	0.04	0.08	0.09	0.02
Elaborative Complexity	1.04	0.005	0.02	0.02	0.05	0.03
Source Count	-0.09	0.005	0.04	0.16	0.16	0.06
Use of Quotes	0.35	0.002	0.03	0.16	0.30	-0.06
Tentativeness	5.54	-0.005	-0.41	0.19	-0.13	-0.24
Focus on the Future	3.12	-0.003	-0.18	-0.28	-0.18	0.16
Focus on the Present	11.87	-0.009	-0.77	0.33	-0.40	-0.18
Focus on the Past	1.92	0.001	--	--	0.06	--
First-Person Singular Pronouns	1.31	-0.002	-0.14	0.46	0.36	-0.04
Third-Person Singular Pronouns	0.74	--	-0.12	-0.01	-0.08	-0.05
Third-Person Plural Pronouns	0.98	--	-0.19	-0.02	-0.15	-0.05
Informal Words	0.34	-0.001	0.06	0.11	0.16	-0.08

We also examined whether linguistic variables could discriminate between a rationale written by a forecaster in a treatment or control group. We selected four comparisons displayed in Table 7. The first comparison was training versus no training; the second, individuals working independently versus in teams; the third, regular team forecasters (with probability training) versus superforecasting teams; the fourth, probability-trained teams accountable to process versus superforecasting teams.

Again, we used LASSO logistic regression to control for Word Count while identifying linguistic-rationale variables that best distinguished whether a rationale came from Group 0 (the control group with rationales marked by dummy variables of 0) or Group 1 (the treatment group with rationales marked by dummy variables of 1). Using the trained model to make in-sample predictions, we measured the power of the model to differentiate group rationales.

Table 7 shows scores for Area Under the Curve (AUC) in the first row. AUC scores can range from .50 (completely random) to 1 (perfectly discriminatory). Table 7 shows that discrimination falls near the lower end of the continuum, but still higher than random guessing.

The model was best at distinguishing trained individuals working independently from trained teams, the highest AUC value (.65). Detecting the differences between trained versus non-trained individuals yields the lowest AUC value (.60).

Table 7 also presents linguistic variables that differentiated group rationales. Intercepts are differences between base rates of group rationales. If Group 1 had more rationales than Group 0, the intercept was positive, otherwise it was negative. First, Word Count had no marginal effects. Second, Comparison Class was most discriminating for training and teaming and less discriminating for superforecasting. Third, Dialectical Complexity was discriminatory of teaming and of superforecasting, and was overall the most discriminatory of the three IC variables. Fourth, Source Count best discriminated the teaming comparison, and Use of Quotes was best at discriminating superforecasters from process-accountable forecasters.

Focusing on the future was somewhat useful for all comparisons, and best at distinguishing superforecasters from process accountable forecasters. Of the three pronoun variables, First-Person Singular was most predictive of teaming and the two superforecasting comparisons. Informal Words, which was positively associated with each of the targeted interventions, was most discriminating of superforecasters from process accountable forecasters.

Table 7. LASSO regression coefficients to isolate effects of specific interventions on each linguistic variable, where N refers to number of rationales from a group. The AUC scores show the power of the model to discriminate between the two groups.

	Group 0: Individuals without training (N = 6,110)	Group 0: Individuals with training (N = 5,409)	Group 0: Teams with training (N = 22,550)	Group 0: Teams with training and process account. (N = 6,377)
	Group 1: Individuals with training (N = 5,409)	Group 1: Teams with training (N = 22,550)	Group 1: Super Teams with training (N = 15,265)	Group 1: Super Teams with training (N = 15,265)
AUC	.60	.65	.64	.62
Variable	LASSO Coefficients			
Intercept	-0.48	0.65	-0.98	0.49
Word Count	--	--	--	--
Comparison Class	1.71	1.77	0.28	--
Integrative Complexity	--	--	0.02	0.06
Dialectical Complexity	--	0.13	0.09	0.11
Elaborative Complexity	0.03	--	0.05	--
Source Count	--	0.57	0.17	0.14
Use of Quotes	--	0.04	0.05	0.08
Tentativeness	-0.01	--	--	0.01
Focus on the Future	-0.01	-0.03	-0.02	-0.05
Focus on the Present	-0.01	0.01	-0.01	-0.01
Focus on the Past	--	-0.02	--	0.01
First-Person Singular Pronouns	-0.03	0.1	0.07	0.08
Third-Person Singular Pronouns	0.01	-0.01	-0.01	--
Third-Person Plural Pronouns	-0.02	--	-0.04	-0.02
Informal Words	0.02	0.06	0.07	0.14

4. Discussion

Across both studies, use of comparison classes (identified by our machine learning model) and integrative complexity (especially dialectical complexity) were most predictive of better forecaster performance. These variables also reliably differentiated between the various pairwise combinations of interventions in Study 2. Focusing on the future (and less on the past) was also a reliable predictor of accuracy.

Other linguistic variables, such as sources, quotes, tentativeness, pronouns, and informality, did not generalize well across studies. Curiously, for instance, a reduction in tentativeness was associated with training, which proved to be an effective intervention for improving accuracy. Perhaps the power of the remaining variables to identify better forecasters depends on tournament ground rules or incentive structure. For example, identifying sources or using quotes may matter less when rationales are visible to all, as doing so might spread information quickly. The impact of teaming (as opposed to working independently) may be more correlated with team dynamics, the use of informal language and social positioning through pronoun choices (Jordan et al., 2019; Kacewicz et al., 2014). In their analysis of top teams, Horowitz et al. (2019) found that team engagement (e.g., references to thinking and collaboration) predicted better team performance. They used topic modeling to identify this factor, whereas we used LIWC variables related to informal language.

5. Conclusion and Future Work

Our results reveal a network of NLP variables correlated with forecasting accuracy, a network consistent with the broad-umbrella "superforecaster" hypothesis (Tetlock & Gardner, 2015). If you want more accurate probability judgments in geopolitical tournaments, you should look for good perspective takers who are tolerant of cognitive dissonance (have high IC and dialectical scores) and who draw adeptly on history to find comparison classes of precedents that put current situations in an outside-view context. Put differently, you should look for forecasters who are exceptions to two of the more robust generalizations of 20th century experimental psychology. Look for forecasters who don't rush to reduce cognitive dissonance (Festinger, 1957) and who don't jump to conclusions from vivid case-specific events (Kahneman, 2011). Effect sizes of these indicators range from small to medium yet were robust across multiple

tournaments. These effects also aligned with our hypotheses about the interventions and how they affect forecasters' thinking patterns. We documented a strong increase in comparison-class usage after training that was informed by Kahneman's (2011) work on heuristics and biases, among other perspectives (Tetlock, 2017). We also observed dialectical complexity increase when forecasters worked in teams versus alone, an effect to be expected if teams encourage perspective taking. Superforecasters showed the highest level of dialectical complexity, an effect that may reflect more dynamic team engagement (explored in detail in Horowitz et al., 2019) as well as individual differences in fluid intelligence and capacity for self-correction as measured by the Cognitive Reflection Test (Mellers et al., 2015).

The dialectical complexity findings also connect to the work of Schwartz et al. (2017) who used traditional NLP metrics to predict (i) how highly a comment was rated by other forecasters for information value, (ii) the impact of the comment on other forecasters' probability judgments; (iii) whether comments helped others form more accurate judgments. Schwartz et al. (2017) found that subordinate conjunctions (e.g. "though", "since", "whereas"), that used complex syntax to connect independent clauses or ideas, were linked to more accurate forecasting updates.

In addition to these empirical contributions, our results provide methodological contributions. We constructed the first NLP classifier for detecting comparison class usage (outside-view thinking) within forecasting rationales and validated the classifier in several ways, including a correlation of $r = .52$ between the model classifier and human-assessed ratings. We found significant positive correlations with forecasting accuracy across multiple studies. The classifier also outperformed users' self-reports of comparison-class usage with higher correlations with performance. Finally, we built on past work on decomposing integrative

complexity into dialectical complexity (evaluative tension between competing perspectives) and elaborative complexity (reinforcing reasons for a forecast) (Conway et al., 2014; Tetlock & Tyler, 1996). Dialectical complexity is a more reliable predictor of forecasting accuracy than elaborative complexity.

Although promising, the current results are but initial steps in a long journey. The NLP correlates are confined to IARPA tournaments and NLP research should be extended to other categories of judgment tasks. Specifically, we recommend exploring the power of NLP indicators to distinguish: (a) linguistic usage among participants in ideological-theoretical debates (e.g., hawks versus doves versus owls debating national security or monetary policy or crime); (b) more from less prescient media commentary on current events, as judged retrospectively with the mixed blessings of hindsight; (c) more from less prescient National Intelligence Estimates (NIE's) and other formal geopolitical assessments (Mandel & Barnes, 2018), again necessarily judged retrospectively. The goal in each case would be to assess how far we can generalize NLP correlates of good judgment in tighly controlled tournaments, with a 100% emphasis on accuracy, to messier real-world situations in which participants make vague-verbiage forecasts that are tricky to assess for accuracy and often blur the distinction between analysis and advocacy. Multi-method validation of NLP indicators across varied institutional settings would be strong evidence that the same habits of mind that serve forecasters well in tournaments also serve political observers struggling to make sense of the news flow in the real world, but working in environments that often elevate persuasion goals over accuracy.

Acknowledgements: We thank the Intelligence Advanced Research Projects Activity (IARPA) for hosting the Global Forecasting Challenge 2 (GFC2) tournament and thank the remaining members of Team KaDSci for their collaboration. We also thank Kathrene Conway and Luke Conway for assistance with the autoIC tool, and Bruce Lawhorn of Basis Technology for use of Rosette software during the GFC2 competition.

Funding:

This research is supported by the Open Philanthropy Foundation as well the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 140D0419C0049. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Armstrong, J. S. (2001). *Principles of forecasting: A handbook for researchers and practitioners*. Kluwer Academic.

Armstrong, J. S. (2005). The forecasting canon: nine generalizations to improve forecast accuracy. *Foresight: The International Journal of Applied Forecasting*, *1*(1), 29–35.

Atanasov, P., Witkowski, J., Ungar, L., Mellers, B., & Tetlock, P. (2020). Small steps to accuracy: Incremental belief updaters are better forecasters. *Organizational Behavior and Human Decision Processes*, *160*, 19–35. https://doi.org/10.1016/j.obhdp.2020.02.001

Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., & Ungar, L. H. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis*, *11*(2), 133–145. https://doi.org/10.1287/deca.2014.0293

Bo, Y. E., Budescu, D. V., Lewis, C., Tetlock, P. E., & Mellers, B. (2017). An IRT forecasting model: Linking proper scoring rules to item response theory. *Judgment and Decision Making*, *12*(2), 90–103.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2

Chang, W., Atanasov, P., Patil, S., Mellers, B. A., & Tetlock, P. E. (2017). Accountability and adaptive performance under uncertainty: A long-term view. *Judgment and Decision Making*, *12*(6), 610–626.

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

Chen, E., Budescu, D. V., Lakshmikanth, S. K., Mellers, B. A., & Tetlock, P. E. (2016). Validating the contribution-weighted model: Robustness and cost-benefit analyses. *Decision Analysis*, *13*(2), 128–152. https://doi.org/10.1287/deca.2016.0329

Conway, L. G., Conway, K. R., Gornick, L. J., & Houck, S. C. (2014). Automated integrative complexity. *Political Psychology*, *35*(5), 603–624. https://doi.org/10.1111/pops.12021

Conway, L. G., Conway, K. R., & Houck, S. C. (2020). Validating automated integrative complexity: Natural language processing and the Donald Trump test. *Journal of Social and Political Psychology*, *8*(2), 504–524. https://doi.org/10.5964/jspp.v8i2.1307

Conway, L. G., Thoemmes, F., Allison, A. M., Towgood, K. H., Wagner, M. J., Davey, K., Salcido, A., Stovall, A. N., Dodds, D. P., Bongard, K., & Conway, K. R. (2008). Two ways to be complex and why they matter: Implications for attitude strength and lying. *Journal of Personality and Social Psychology*, *95*(5), 1029–1044. https://doi.org/10.1037/a0013336

Cross, D., Ramos, J., Mellers, B., Tetlock, P. E., & Scott, D. W. (2018). Robust forecast aggregation: Fourier L2E regression. *Journal of Forecasting*, *37*(3), 259–268. https://doi.org/10.1002/for.2489

Friedman, J. A., Baker, J. D., Mellers, B. A., Tetlock, P. E., & Zeckhauser, R. (2018). The value of precision in probability assessment: Evidence from a large-scale geopolitical forecasting tournament. *International Studies Quarterly*, *62*(2), 410–422. https://doi.org/10.1093/isq/sqx078

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. https://doi.org/10.1198/016214506000001437

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statical learning: Data mining, inference, and prediction*. Springer. https://doi.org/10.1198/jasa.2004.s339

Horowitz, M., Stewart, B. M., Tingley, D., Bishop, M., Samotin, L. R., Roberts, M., Chang, W., Mellers, B., & Tetlock, P. (2019). What makes foreign policy teams tick: Explaining variation in group performance at geopolitical forecasting. *Journal of Politics*. https://doi.org/10.1086/704437

Houck, S. C., Conway, L. G., & Gornick, L. J. (2014). Automated integrative complexity: Current challenges and future directions. *Political Psychology*, *35*(5), 647–659. https://doi.org/10.1111/pops.12209

Jervis, R. (1997). *System effects: Complexity in political and social life*. Princeton University Press.

Jordan, K. N., Sterling, J., Pennebaker, J. W., & Boyd, R. L. (2019). Examining long-term trends in politics and culture through language of political leaders and cultural institutions. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(9), 3476–3481. https://doi.org/10.1073/pnas.1811987116

Jose, V. R. R., Nau, R. F., & Winkler, R. L. (2009). Sensitivity to distance and baseline distributions in forecast evaluation. *Management Science*, *55*(4), 582–590. https://doi.org/10.1287/mnsc.1080.0955

Kacewicz, E., Pennebaker, J. W., Davis, M., Jeon, M., & Graesser, A. C. (2014). Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, *33*(2), 125–143. https://doi.org/10.1177/0261927X13502654

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus, and Giroux.

Lee, I. A., & Preacher, K. J. (2013). *Calculation for the test of the difference between two dependent correlations with one variable in common [Computer software]*. http://quantpsy.org

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, *125*(2), 255–275. https://doi.org/10.1037//0033-2909.125.2.255

Lustick, I. S., & Tetlock, P. E. (2020). The simulation manifesto: The limits of techno-empiricism in geopolitical forecasting. *Futures & Foresight Science*.

Mandel, D. R., & Barnes, A. (2018). Geopolitical forecasting skill in strategic intelligence. *Journal of Behavioral Decision Making*, *31*(1), 127–137. https://doi.org/10.1002/bdm.2055

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

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

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 a geopolitical forecasting tournament. *Psychological Science*, *25*(5), 1106–1115. https://doi.org/10.1177/0956797614524255

Merkle, E. C., & Steyvers, M. (2013). Choosing a strictly proper scoring rule. *Decision Analysis*, *10*(4), 292–304. https://doi.org/10.1287/deca.2013.0280

Moore, D. A., Swift, S. A., Minster, A., Mellers, B., Ungar, L., Tetlock, P., Yang, H. H. J., & Tenney, E. R. (2017). Confidence calibration in a multiyear geopolitical forecasting competition. *Management Science*, *63*(11), 3552–3565. https://doi.org/10.1287/mnsc.2016.2525

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. https://doi.org/10.1037/0033-295X.84.3.231

Page, S. E. (2018). *The model thinker: What you need to know to make data work for you*. Basic Books.

Pennebaker, J. W., Booth, R. J., Boyd, R. L., & Francis, M. E. (2015). *Linguistic inquiry and word count: LIWC2015*. Pennebaker Conglomerates.

Raleigh, C., Linke, A., Hegre, H., & Karlsen, J. (2010). Introducing ACLED: An armed conflict location and event dataset. *Journal of Peace Research*, *47*(5), 651–660. https://doi.org/10.1177/0022343310378914

Satopää, V. A., Baron, J., Foster, D. P., Mellers, B. A., Tetlock, P. E., & Ungar, L. H. (2014). Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting*, *30*(2), 344–356. https://doi.org/10.1016/j.ijforecast.2013.09.009

Satopää, V. A., Salikhov, M., Tetlock, P., & Mellers, B. (2020). Bias, information, noise: The BIN model of forecasting. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3540864

Schroder, H. M., Driver, M. J., & Streufert, S. (1967). *Human information processing: Individuals and groups functioning in complex social situations*. Holt, Rinehart and Winston.

Schwartz, H. A., Rouhizadeh, M., Bishop, M., Tetlock, P., Mellers, B., & Ungar, L. H. (2017). Assessing objective recommendation quality through political forecasting. *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2348–2357. https://doi.org/10.18653/v1/d17-1250

Scoblic, P., & Tetlock, P. E. (2020). Beyond crystal balls: A better way to plan. *Foreign Affairs*, *99*, 10.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*(2), 245–251. https://doi.org/10.1037/0033-2909.87.2.245

Suedfeld, P., & Tetlock, P. (1977). Integrative Complexity of Communications in International Crises. *Journal of Conflict Resolution*, *21*(1), 169–184. https://doi.org/10.1177/002200277702100108

Suedfeld, P., & Tetlock, P. E. (2001). Individual differences in information processing. In *Blackwell international handbook of social psychology: Intra-individual processes*. Blackwell Publishers.

Suedfeld, P., & Tetlock, P. E. (2014). Integrative complexity at forty: Steps toward resolving the scoring dilemma. *Political Psychology*, *35*(5), 597–601. https://doi.org/10.1111/pops.12206

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Tetlock, P. E. (2017). *Expert political judgment*. Princeton University Press. https://doi.org/10.1515/9781400888818

Tetlock, P. E., & Gardner, D. (2015). *Superforecasting: The art and science of prediction*. Crown Publishers.

Tetlock, P. E., & Kim, J. I. (1987). Accountability and judgment processes in a personality prediction task. *Journal of Personality and Social Psychology*, *52*(4), 700–709. https://doi.org/10.1037//0022-3514.52.4.700

Tetlock, P. E., Mellers, B. A., Rohrbaugh, N., & Chen, E. (2014). Forecasting tournaments: Tools for increasing transparency and improving the quality of debate. *Current Directions in Psychological Science*, *23*(4), 290–295. https://doi.org/10.1177/0963721414534257

Tetlock, P. E., & Tyler, A. (1996). Churchill's cognitive and rhetorical style: The debates over nazi intentions and self-government for India. *Political Psychology*, *17*(1), 149. https://doi.org/10.2307/3791947

Yaniv, I., Yates, J. F., & Smith, J. E. K. (1991). Measures of Discrimination Skill in Probabilistic Judgment. *Psychological Bulletin*, *110*(3), 611–617. https://doi.org/10.1037/0033-2909.110.3.611

Zong, S., Ritter, A., & Hovy, E. (2020). Measuring forecasting skill from text. In *arXiv*. arXiv. https://doi.org/10.18653/v1/2020.acl-main.473

Appendix A.

The CHAMPS KNOW guidelines are rooted in a variety of scholarly traditions. The first six components ("CHAMPS") are grounded in Bayesian probability theory and work on bounded rationality, and cognitive biases and debiasing. The last four ("KNOW") are grounded in an eclectic mix of theories of political behavior, including applied game theory, institutionalism and work on dynamic systems. The CHAMPS KNOW guidelines are as follows:

- C - Comparison classes. Provide relevant classes and calculate base-rates.
- H - Hunt for info. Share the evidence that helps inform your prediction.
- A – Adjust and update. Explain the reason for updating your prediction.
- M - Models. Explain any application of mathematical or statistical models.
- P - Post-mortem. Interpret past failures and successes, draw lessons.
- S - Select the right questions to answer.
- K - Know the power-players. Describe individuals with influence and how will they use it.
- N - Norms & protocols of institutions. Identify laws and rules that matter.
- - Other perspectives. Catch blind spots (revolutions, technologies, etc.) that upset expectations.
- W – Wildcards and black swans. Reflect on the border between known and unknown unknowns.

Appendix B.

The seven levels of IC are:

1. One-dimensional, categorical, right/wrong reasoning
2. Implicit acknowledgment of differing views
3. At least two perspectives explicitly developed
4. Connections between perspectives implied
5. First-order connections explicitly developed
6. High order connections among perspectives implicit
7. Explicit integration of complex interactions or trade-offs

As an example, consider the following rationales for the question of "Will North Korea launch a medium-range or longer ballistic missile between 20 June 2019 and 30 September 2019?":

(1) "North Korea launched a short-range ballistic missile and two types of rockets on two occasions earlier in May. North Korea has no fear when it comes to testing their missiles. I see no slow down in sight and they will likely test again in this time period if for no other reason than to show their independence and to tick off those who oppose these tests."

(2) "North Korea acts in a predictable pattern - sanctions, act out, demand assistance usually food, get aid - then back to sanctions. Right now, North Korea has been playing along with Trump's efforts for personal diplomacy, making empty promises and appeasing the president. However, now he is distracted among many different issues and crises, from his multiple trade disputes, to Iran a more serious threat than NK, to his upcoming re-election. North Korea may feel forgotten about and need to do something to raise their international profile. Based off that link, NK was doing multiple medium range missile tests in 2017, which brought Trump to the

table to negotiate. Now, two years later, it is not impossible they will return to such tactics. However, too much attention could get them in trouble with China. So while possible, this is not exceedingly likely."

The first rationale was rated 1.5 on IC using the autoIC calculator, whereas the second, longer, more detailed, gets a score of 4.

Appendix C.

We sampled 100 Questions from the set of 399, and for each, randomly chose one rationale that had a comparison class calculation, and another that did not. The selection of two rationales per question reduced the risk of biasing a machine-learning model. If all rationales that featured comparison classes were from questions related to commodity prices, whereas non-comparison class forecasts were related to the more idiosyncratic questions (e.g., Brexit), the algorithm might associate the words "price", "gold", or "crude" as comparison class identifiers, regardless of whether the rationale contained a comparison class.

With the 100 pairs of forecasts, we labeled rationales with a comparison class as 1.0 and non- comparison class rationales as 0.0. Then for varying sample sizes {5, 10, 20, 40, 60}, we randomly selected N questions (2*N rationales) as a training sample, and converted the rationales to a Document Term Frequency (DTF) matrix[9] which featured the most frequently used terms. In this way, the more prominent words are extracted by the machine. We then fit a Random Forest (RF) model to the DTF matrix data and made probabilistic predictions for the 2*(100-N) rationales in terms of whether each rationale was labeled as base-rate. For each value of N in {5,10,20,40,60}, we repeated the sampling of N questions 100 times, each time calculating an Area Under the Curve (AUC) score for the questions not within the training sample to gauge the power of the RF prediction to distinguish comparison class from non-comparison class rationales. Figure A1 shows the averaged AUC (over 100 iterations) as a function of training sample size for the RF models fit. As the training sample grows, so does predictive performance.

[9] Which is a fancy way of saying we created a matrix that featured the most frequently used words in the corpus as columns, with the first row denoting the count of such words that appeared in first rationale, the second row denoting the count of such words in the, second rationale, and so on.

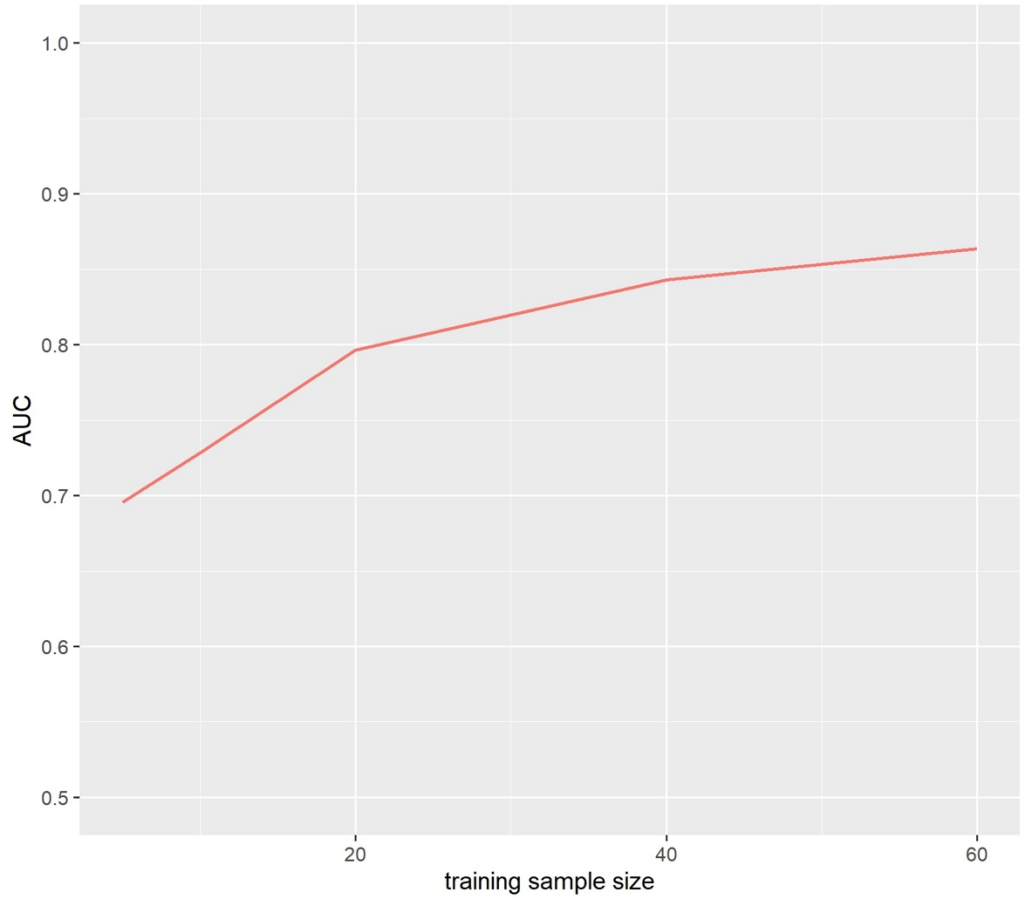

Figure A1. AUC scores for RF predictor of comparison class usage for training samples of various sizes.

Using the entire set of 200 rationales, we then fit a final RF model to predict comparison class usage on all rationales, with such prediction denoted as *Comparison Class*, with $0 \leq$ *Comparison Class* ≤ 1. As a final validity test, we sampled over 300 additional rationales from the GFC2 data and one author independently and blindly graded them on a Likert-style scale of 0 (no comparison class) to 3 (a comparison class clearly used). We then used the RF model to assign a *Comparison Class* prediction to each rationale. The correlation between the *Comparison Class* predicted values and Likert ratings was $r(332) = .52$, a large effect size, justifying confidence that we can automate identification of comparison classes.