Judgment and Decision Making, Vol. 7, No. 6, November 2012, pp. 728–740

Using inferred probabilities to measure the accuracy of imprecise forecasts

Paul Lehner* Avra Michelson† Leonard Adelman‡ Anna Goodman†

Abstract

Research on forecasting is effectively limited to forecasts that are expressed with clarity; which is to say that the forecasted event must be sufficiently well-defined so that it can be clearly resolved whether or not the event occurred and forecasts certainties are expressed as quantitative probabilities. When forecasts are expressed with clarity, then quantitative measures (scoring rules, calibration, discrimination, etc.) can be used to measure forecast accuracy, which in turn can be used to measure the comparative accuracy of different forecasting methods. Unfortunately most real world forecasts are not expressed clearly. This lack of clarity extends to both the description of the forecast event and to the use of vague language to express forecast certainty. It is thus difficult to assess the accuracy of most real world forecasts, and consequently the accuracy the methods used to generate real world forecasts. This paper addresses this deficiency by presenting an approach to measuring the accuracy of imprecise real world forecasts using the same quantitative metrics routinely used to measure the accuracy of well-defined forecasts. To demonstrate applicability, the *inferred probability method* is applied to measure the accuracy of forecasts in fourteen documents examining complex political domains.

Key words: inferred probability, imputed probability, judgment-based forecasting, forecast accuracy, imprecise forecasts, political forecasting, verbal probability, probability calibration.

1 Introduction

Forecasting accuracy, and the determination of practices, methods and tools that improve accuracy, is a topic of substantial research and practical importance. (See Armstrong, 2001, and Tetlock, 2005, for introductions.) When endeavoring to measure forecast accuracy, researchers generally require that the forecasted events be clearly described and that the degree of forecast certainty be expressed as quantitative probabilities.

In contrast to forecasting research, most published forecasts describe forecast events with considerable imprecision and use vague certainty expressions (Gardner, 2010). This is particularly true of forecasts about complex international political events, which is a substantive domain of interest to us. Consider for example the statement from the Stratfor[1] forecasts for 2006 for Iran:

> Ayatollah Mesbah Yazdi has a fair chance of making it into the Assembly of Experts when elections take place.

It's rather difficult to gauge what is meant by "fair chance". Is that a 50% chance, a 20% chance, or perhaps an 80% chance? If the event occurs, it would be unclear if it should be judged as a mostly accurate or inaccurate forecast.

Consider also the following statement from the Stratfor Iran 2006 forecast:

> The Iranian nuclear program crisis likely will result in Tehran eventually backing down.

In addition to the fact that different individuals vary widely in their interpretation of the word "likely" (e.g., from 0.3 to 0.8 in Beyth-Marom, 1983), the phrase "backing down" is itself hard to define. If there is a negotiated settlement does that mean that Iran "backed down", that the other parties "backed down", or that a settlement was found where everyone could claim success? In such cases it is difficult to specify clear criteria *a priori* for determining whether or not the event occurred.

Anecdotally, authors of forecasting documents have expressed to us vigorous arguments in favor of imprecise forecasts. They believe that clarity requirements severely limit their ability to express what they intend to say. For example, a phrase such as "backing down" succinctly describes an important element of a forecast event—namely that the individuals involved will be accepting an option that is less desirable than their expressed preference. In addition many forecasters prefer verbal certainty expres-

This research was funded by the MITRE Corporation under IR&D project number 05MSR003-FA.

*The MITRE Corporation, 7594 Colshire Drive, McLean, Virginia 22102-7539. Email: plehner@mitre.org.

†The MITRE Corporation

‡George Mason University

[1] Stratfor is a public company that "Provides strategic intelligence on global business, economic, security and geopolitical affairs." Many of the forecast statements we examine in this study are found on the Stratfor website, www.stratfor.com, which is available through a paid subscription service.

sions to quantitative uncertainties because they believe the later connotes artificial precision and misleads readers.

Though there may be valid reasons for expressing forecasts imprecisely there is still a need to evaluate forecast accuracy. Research in expert political judgment (e.g., Tetlock, 2005) would suggest that, lacking objective feedback on the accuracy of their forecasts, experts are unlikely to accurately determine when their forecasts were inaccurate. In particular, given the hindsight memory bias, we would anticipate that forecasters will remember certainty statements such as "fair chance" as having meant a low probability when the event didn't occur and a high probability when the event did occur. Furthermore, without measuring the accuracy of real world forecasts it is difficult to compare the accuracy of different forecasting methods outside of artificial research settings—and unfortunately results from artificial research settings are widely ignored by practitioners.

The distinction between academic research and practice is well illustrated by the fourteen years of research with political analysts summarized in Tetlock (2005). Every research effort described by Tetlock involved constructing carefully defined forecasting questions and asking analysts to provide quantitative forecasts. The forecasts that the analysts actually published were not examined.

In this paper we describe a method for measuring the accuracy of imprecise forecasts. Our objective is to add sufficient rigor to the evaluation of imprecise forecasts to enable the application of common metrics of forecast accuracy to forecasts that are actually published.

Our approach incorporates two basic techniques: inferred probabilities and impartial ground truth judgments. First we use *inferred probabilities* to impute quantitative probabilities from verbal expressions of certainty. Simply put, we ask multiple readers to assign quantitative probabilities based on their understanding of the written document; rather than their personal beliefs. Second we ask multiple *ground truth raters*, who do not see the original documents or inferred probabilities, to independently research and estimate whether or not a forecasted event has occurred. In addition, as needed, we use inter rater agreement data to statistically adjust estimated ground truth frequencies.

Below we present the details of our current instantiation of the inferred probability method. We note that this is just one of multiple possible instantiations of this general approach, and that researchers should tailor the inferred probability method approach to the specific needs of their research program.

2 The Inferred Probability Method

Below are the basic steps of the inferred probability method for measuring the accuracy of published forecasts, no matter how imprecisely those forecast events and certainties are worded.

1. *Extract forecast events.* Specific possible future events that are referenced in the forecast document are identified and extracted. An explicit and repeatable protocol is used to for identifying these forecast events.[2] As used in the inferred probability method, this protocol extracts events that are expressed without condition and does not include forecast events that are conditioned on other events.
2. *Infer probabilities.* The list of forecasted events, along with the original forecast document, is given to multiple readers. These readers are asked write down their inferred probability for each event. In some cases the period for the forecast is not clearly identified in the forecast document. In such cases we specify a time period and ask readers to infer probabilities for that time period. Individual reader inferred probabilities are averaged.
3. *Impartial estimate of ground truth.* For each forecast event, the event described in each forecast statement is listed in a separate table along with the time period for which the forecast event applies. Any indication of whether or not the original forecast document indicated that the event would or would not occur is removed. Ground truth raters, who are either subject matter experts (SMEs) or document researchers[3] and who do not see the original forecast statement, are then asked to retrospectively evaluate whether or not each event occurred. Each ground truth rater works independently. They use the following scale:

   (a) True (the event occurred)
   (b) Not sure, but I tend to think that it occurred
   (c) I don't know [4]
   (d) Not sure, but I tend to think that it did not occur
   (e) False (the event did not occur)

   Using one of several methods, individual ratings are then combined into a ground truth assignment for each forecast event.

[2] The full protocol is among the addendum items listed in the Table of Contents of the journal.

[3] A "document researcher" is someone who examines relevant publicly available documents (e.g., news sources) to find requested information. Their professional expertise is in information research rather than topic knowledge.

[4] This can be used by ground truth raters if either they truly don't know or if the event statement is too vague to determine ground truth no matter how much they know.

4. *Estimate Accuracy.* Once ground truth estimates are received, then accuracy is measured using exactly the same measures and procedures used to estimate the accuracy of other quantitative probability forecasts.

To illustrate these steps, and describe some of the important nuances of our instantiation of the inferred probability method, consider the following statements selected from the declassified US National Intelligence Estimate entitled *Prospects for Iraq's Stability: A Challenging Road Ahead* (2007).

> ... Arab groups in Kirkuk continue to resist violently what they see as Kurdish encroachment
>
> ... Iraq's neighbors influence, and are influenced by, events within Iraq, but the involvement of these outside actors is not likely to be a major driver of violence or the prospects for stability because of the self-sustaining character of Iraq's internal sectarian dynamics.
>
> ... Syria continues to provide safe haven for expatriate Iraqi Bathists and to take less than adequate measures to stop the flow of foreign jihadists into Iraq.
>
> ... Turkey does not want Iraq to disintegrate and is determined to eliminate the safe haven in northern Iraq of the Kurdistan People's.

Table 1, which includes actual data from our study, illustrates the results of each step of this process. The first column in Table 1 shows the results of Step 1. The different events included in the overall forecast statement are separately listed. Note that we included event statements where resolving ground truth would obviously be very difficult. For example, the statement "The involvement of outside actors will not be a major driver of violence in Iraq" would seem to be difficult to resolve simply because the expression "major driver" is not well defined. Rather than exclude such statements, we included as many of the forecast events as possible, even if they appeared to be difficult to resolve. Furthermore our protocol tries to ensure that the wording of each forecast event statement corresponds closely to the original wording in the forecast document, so we did not change the event statements in a way that would have made them easier to resolve. We left it to the readers to decide if event statements were sufficiently well defined to infer a probability and to the ground truth raters to decide if it the events were resolvable.

The second column shows the results of the second step, where three different readers were asked to infer a probability for each forecast event from the forecast document. Although different readers had somewhat different inferred probabilities, those inferred probabilities were well correlated. As illustrated by Reader 3, who did not answer two of inferred probability questions, all readers had the option of not answering and they did so for a variety of reasons (multiple interpretations of the event statement, they believed the event was conditioned on a hypothetical event, etc.).

The third column shows the results of the third step. Two sets of independent ground truth judgments were recorded about whether each of the statements in the first column were true or false. The ground truth raters saw only the individual forecast events listed in the first column (converted to past tense), but did not see the original forecast document or the inferred probabilities. Consequently the raters were not told anything about the original forecast and did not know if the event was forecasted to occur or to not occur. As long as one rater assigned either a "Yes" or "No" rating, and the other rater didn't disagree, then we accepted that "Yes" or "No" rating as ground truth.[5] (Later in the paper we measure the implications of this weak criterion for assigning ground truth.)

Regarding Step 4 the fourth column shows the base error score. Mean Absolute Error (MAE) for these six forecasts is 0.33.[6] There are too few data points in this example to show a calibration curve, but for illustration note that for the five forecasts that clustered around 0.85, four of those events where judged to have occurred. So the initial estimate for 0.85 certainty judgments is that 80% of those events occurred.

Finally we can employ a procedure to convert the level of inter-rater agreement on the ground truth into an estimate of the accuracy of ground truth ratings which, in turn, we can use to estimate ground truth probabilities and statistically adjust the accuracy measures. This statistical adjustment is not essential to our method, so it is described near the end of this paper.

3 Applying the Inferred Probability Method to 14 documents

We applied the steps described above to measure the forecast accuracy of 14 forecast documents. There were two

[5] Our data analysis ignored the fact that raters had a graded 5 point scale for assigning ground truth. We treat a rating of "5-True (the event occurred)" and "4-Not sure, but I tend to think that it occurred" as simply a True rating. We provided raters the 5 point scale simply to encourage them to assign more True and False ratings.

[6] While there are many individual exceptions, forecasting researchers often prefer to use mean absolute error, researchers examining human judgment and decision making often use quadratic error scores, and researchers with an interest in Bayesian reasoning and inference prefer a log error score. In this paper we use mean absolute error because it is easily understood. We do not claim that absolute error is in any sense a proper scoring rule.

Table 1: Results of inferred probability evaluation process for a few forecast events

Forecast event	Inferred probabilities	Ground truth ratings	Base error score
Arab groups in Kirkuk will resist violently what they see as Kurdish encroachment in the January 2007 to July 2009 time frame.	Reader1 = 0.90 Reader2 = 0.90 Reader3 = 0.85 Average = **0.88**	Yes, Unk	0.12
The involvement of outside actors will not be a major driver of violence in Iraq in the January 2007 to July 2009 time frame.	Reader1 = 0.80 Reader2 = 0.85 Reader3 = na Average = **0.83**	No, Unk	0.83
The involvement of outside actors will not be a major driver of stability in Iraq in the January 2007 to July 2009 time frame.	Reader1 = 0.80 Reader2 = 0.85 Reader3 = na Average = **0.83**	Yes, Unk	0.17
Syria will provide a safe haven for expatriate Iraqi Bathists in the January 2007-June 2009 time frame.	Reader1 = 0.90 Reader2 = 0.95 Reader3 = 0.70 Average = **0.85**	Yes, Yes	0.15
Syria will take less than adequate measures to stop the flow of foreign jihadists into Iraq in the January 2007-June 2009 time frame.	Reader1 = 0.90 Reader2 = 0.95 Reader3 = 0.70 Average = **0.85**	Yes, Yes	0.15
Turkey will eliminate the safe haven in northern Iraq of the Kurdistan People's Congress in the January 2007 to July 2009 time frame.	Reader1 = 0.70 Reader2 = 0.30 Reader3 = 0.50 Average = **0.50**	Yes, No	n/a

purposes for this test application. First we wanted to gauge whether the inferred probability method could reasonably assess the accuracy of imprecise forecasts. There were many points of possible failure. The readers' inferred probabilities may be so divergent as to make it difficult to claim that the documents forecast anything. The ground truth raters may find the event statements too imprecise to even rate. Accuracy results may be very divergent from research on forecast accuracy suggesting that the measurements are not comparable. Second, we were specifically interested in the accuracy profile of these documents. This paper concentrates on the first purpose of this study—to evaluate the inferred probability method itself. Substantive implications of this study are examined elsewhere.

3.1 Materials

We applied the steps described above to measure the forecast accuracy of fourteen documents. This included nine documents produced by Stratfor; three documents from Jane's and the declassified key judgments section of two National Intelligence Estimates (NIEs).

As noted above, Stratfor is a public company that provides intelligence for various consumers. Jane's is well known for its work in documenting worldwide military capabilities, but it also provides some analytic documents with specific sections entitled "forecasts".[7] Finally the NIEs are considered to be the premier analysis product of the US Intelligence Community. NIEs reflect the aggregate judgment of multiple intelligence organizations on key topics. The word "estimate" is often used in the intelligence community as a euphemism for judgment-based forecasts.

The following are the specific documents we examined, the time periods when readers inferred probabilities and when ground truth ratings occurred.

[7]The Jane's documents can be found at www.janes.com. Like Stratfor, a paid subscription is required to access these documents.

Group 1: (Inferred February 2010, resolved April 2010)
Jane's 2006 Forecast for Iran ("Further JID Forecasts," 2006)
Stratfor 2006 Forecast for Iran ("Middle East," 2006)
Stratfor 2006 Forecast for South Africa ("Sub-Saharan," 2006)
Stratfor 2006 Forecast for Sudan ("Sub-Saharan," 2006)

Group 2: (Inferred April 2010, resolved June 2010)
Jane's: *US and Iran: Road Map to Conflict* (2007)
Jane's: *Larijani's Departure Fuels Iran Power Struggle* (2007)
Stratfor 2007 Forecast for Iran ("Middle East," 2007)
Stratfor 2007 Forecast for South Africa ("Africa," 2007)
Stratfor 2007 Forecast for Sudan ("Africa," 2007)

Group 3: (Inferred August 2010, resolved November 2010)
NIE *Prospects for Iraq Stability* (2007)
NIE *Trends in Global Terrorism* (2006)

Group 4: (Inferred March 2011, resolved February 2012)
Stratfor 2011 Forecast for Iran and Iraq ("Middle East," 2011)
Stratfor 2011 Forecast for South Africa ("Sub-Saharan," 2011)
Stratfor 2011 Forecast for Sudan ("Sub-Saharan," 2011)

The first three groups of documents were selected in part because they partially overlap the countries and time periods of a study described by Mandel, Barnes and Hannigan (2009) where analysts directly expressed forecast certainties as quantitative probabilities. They provide a possible standard against which to evaluate our inferred probability method. The fourth group was selected to cover the same topic areas, but allowed us to examine documents where probabilities were inferred prospectively, before the events were supposed to occur.

Six members of the MITRE Corporation were asked to read and infer probabilities for the documents. Three of the readers inferred probabilities for all fourteen documents. Two of these readers had more than five years professional experience in intelligence analysis and the third had several decades of professional national policy experience. The two NIEs in Group 3 were reviewed by these same three readers plus an additional three readers; one with more than five years of professional intelligence analysis experience, the second with more than five years of policy experience and the third with more than five years of professional legal experience.

Ground truth was assessed by a mixture of SMEs and document researchers. At least one SME (subject matter expert) assessed ground truth for each document.

3.2 Procedures

The procedure described above for extracting forecast events and evaluating ground truth was applied. However after each group we re-examined our procedure and made some minor procedural changes. For Group 1 we asked the readers only to infer probabilities, and two raters to assign ground truth, for the time period specified in the forecast documents. Some of the documents in Groups 2 and 3 did not have a clear forecast time period, so we specified one. Finally, for Group 4, after providing their inferred probabilities, we also asked the readers to provide their personal probabilities for each forecast event.

4 Results

We first summarize the results for all 14 documents. We then examine and compare various document subsets to examine several hypotheses related to the viability of the inferred probability method.

Across these 14 documents there were 237 forecast events. For the most part readers had little difficulty in assigning inferred probabilities. In the few cases where they did have difficulty, it was often because they had more than one possible interpretation of the event statement. For all but one of the 237 events, the majority of readers assigned an inferred probability. There were three readers who read all 14 documents. For the 201 events where all three readers inferred a probability, the intra-class correlation was 0.702 (model 2, individual). We note that there were statistically significant differences between readers where for some readers the interpreted probability was on average higher than for other readers ($p<.0001$). The greatest difference was between two readers with average inferred probabilities of 0.711 and 0.619. These three readers plus an additional three read the two NIEs. For 50 of the 71 events in these documents all six readers inferred a probability, and for these 50 the intra class correlation was 0.623; again with significant difference between the readers.

Across the 237 forecast events, whenever two ground truth raters assigned either a True or False ground truth rating, inter rater agreement was 79%. The majority rule method was used to assign ground truth. This resulted in 115 true events and 72 false events. Another method for assigning ground truth (described later) that estimated ground truth probabilities and applied an 85% certainty threshold yielded exactly the same ground truth assignments.

4.1 Accuracy profile

Table 2 shows, for each document, the average inferred probability for events that did and did not occur, as well

Table 2: Average probabilities and mean absolute error (MAE) for 14 forecast documents.

	Average probability for Events that did Occur (n)	Average probability for Events that did not Occur (n)	Mean Absolute Error (MAE)
NIE 2006 Prospects for Iraq Stability	.622 (12)	.578 (14)	0.486
NIE 2006 Trends in Global Terrorism	.785 (29)	.538 (8)	0.285
Jane's 2006 Forecast for Iran	.817 (3)	.433 (3)	0.308
Jane's: US and Iran: Road Map to Conflict (Feb 2007)	.744 (3)	n/a (0)	0.256
Jane's: Larijani's Departure Fuels Iran Power Struggle (Nov 2007)	.725 (2)	.700 (2)	0.488
Stratfor 2006 Forecast for Iran	.763 (8)	.604 (4)	0.359
Stratfor 2007 Forecast for Iran	.803 (10)	.568 (9)	0.373
Stratfor 2011 Forecast for Iran and Iraq	.323 (20)	.496 (8)	0.625
Stratfor 2006 Forecast for South Africa	.717 (3)	.883 (1)	0.433
Stratfor 2007 Forecast for South Africa	.722 (6)	.794 (3)	0.450
Stratfor 2011 Forecast for South Africa	.857 (7)	.900 (1)	0.238
Stratfor 2006 Forecast for Sudan	.758 (2)	.672 (3)	0.500
Stratfor 2007 Forecast for Sudan	.850 (8)	.892 (2)	0.298
Stratfor 2011 Forecast for Sudan	.737 (2)	.075 (14)	0.099
All Forecasts	**.744 (115)**	**.517 (72)**	**.356**

as the mean absolute error. The absolute difference between the inferred probabilities of events that occurred and events that did not occur was only 0.227. Of particular note is the fact that the average probability of events that did not occur was above 0.5. Mean Absolute Error (MAE) was 0.356; which is quite high when one considers that a MAE of 0.50 can be achieved by always asserting 0.5 for all forecasts. These results suggest very poor accuracy, but the picture is very different when one examines calibration.

Figure 1 shows the calibration curve for the combined list of 187 forecast events. Each probability level is composed of forecasted events where the average inferred probability rounded to that probability level. So, for example, if three readers had inferred probabilities of 0.9, 0.95 and 0.79, then the average inferred probability is 0.88 and that event forecast would appear at the 0.9 level.

On average, the absolute difference between the observed relative frequency and perfect calibration was 0.11.[8] We note further that there is a negative correlation between sample size and absolute difference from perfect calibration (-0.31, n.s.). That is to say, probability levels with larger sample sizes exhibited better calibration. This suggests that, if more data were collected, calibration might improve.

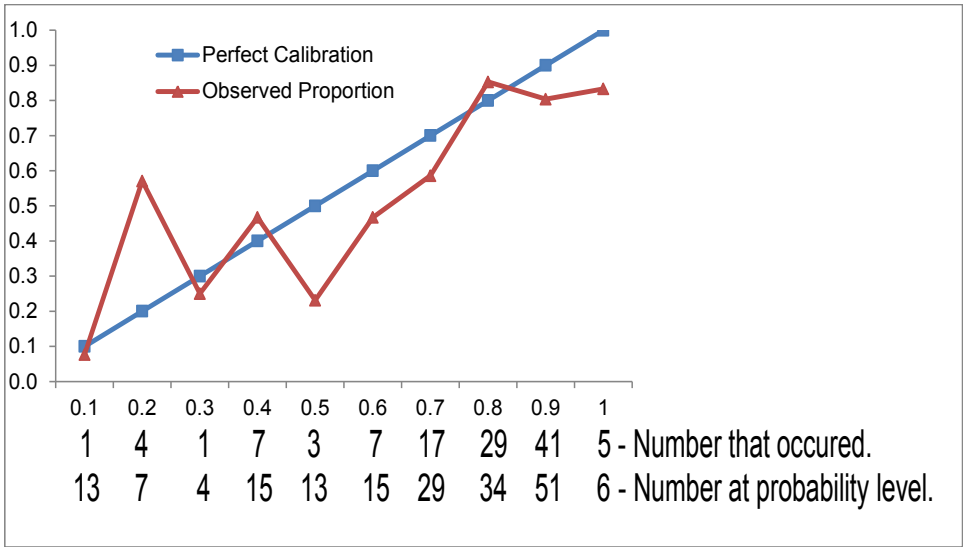
So the discrimination[9] and error score measures suggest very poor accuracy while the calibration measure suggests good accuracy. The explanation for this discrepancy appears to be two *reporting biases* in the forecasts that the authors select to include in these documents.

The first reporting bias is a tendency to avoid including obvious forecasts in these documents; that is to say forecasts with probabilities close to 0.0 or 1.0. This reporting bias is a consequence of two factors: topic selection and space consideration. First, these documents address complex political topics of international significance. In such topic areas it might be expected that very few forecasts would be "easy calls" with probabilities close to 0.0 or 1.0. Second, these documents are short documents intended as summative readings for policy and decision makers. In such documents authors will presumably tend to avoid wasting space on obvious forecasts and instead will write about the more perplexing issues. Consequently, both topic selection and a tendency

[8] This is a weighted average based on the number of observation at each level. The unweighted average is 0.134.

[9] Above we measured "discrimination" by taking the difference between the means. There are other measures of discrimination such as a normalized difference between means (Wallsten, Budescu & Zwick, 1997) and a measure called the discrimination index (Yaniv, Yates & Smith, 1991) that is described below. Our comments on discrimination apply to all of these measures.

Figure 1: Calibration curve for the combined NIE, Jane's and Stratfor forecast events.



to avoid forecasting the obvious will ensure that the bulk of the forecasts are between 0.1 and 0.9; with most being between 0.2 and 0.8. This distribution of probability estimates ensures a relatively high error score no matter what the outcomes.

The second reporting bias is a tendency of authors to describe the future in terms of events that are likely to occur rather than by stating events that are unlikely to occur. Furthermore when unlikely events were discussed, they were on occasion phrased in a way to suggest that it was likely that the event would not occur. Since our protocol for extracting event statements maintained the original language in the document, the inferred probabilities for these events were above 0.5. Consequently both event selection and writing style ensured that the bulk of the inferred probabilities were above 0.5. This bias ensures that discrimination measures will show poor results no matter what the outcomes.

Because of these reporting biases, we believe that error scores and discrimination measures are largely uninformative accuracy measures when applied to forecast documents. The situation would be analogous to asking a forecaster to assign probabilities to 100 events, but to reveal the probability for only 10 events with probabilities close to 0.75. If the forecaster is perfectly calibrated, then these instructions ensure an expected mean absolute error of 0.37 or worse, and discrimination close to 0.0, no matter what the outcomes.

Calibration is entirely different. These documents exhibited good calibration in their forecasts and we can think of no reason why this result would be an artificial consequence of the reporting biases. Indeed, since these documents tend to include the most perplexing topics and forecasts, it seems reasonable to believe that measured calibration would be slightly better if the more obvious and easy forecasts were included.

Overall these results present a clear profile of the accuracy of these documents.

1. These documents generally included forecasts for important events that the authors believed had a greater than 0.5 probability of occurrence. Consequently, even for forecast events that did not occur, the average inferred probability was greater than 0.5.
2. The expressions of forecast certainty appear somewhat conservative. When converted to inferred probabilities, forecast certainties are rarely interpreted as definitive (i.e., inferred probability is 1.0 or 0.0).
3. The inferred probabilities are reasonably well calibrated.

As mentioned above, this paper examines the viability of the inferred probability method as a method for quantitatively measuring the accuracy of forecasts expressed with imprecision. Our intent is for this method to be used to assess the accuracy of numerous significant forecasts that are expressed with imprecision and to further assess the comparative accuracy of the methods and tradecraft that resulted in those forecasts.

Although the above results clearly demonstrate that the method is executable; that does not imply that the results are meaningful or useful. Below we consider three important issues related to the meaningfulness of the measures yielded by the inferred probability method. Namely,

1. Whether the accuracy results are commensurate to accuracy results obtained by directly querying analysts for quantitative probabilities.
2. Whether the quantitative accuracy results are affected by reader personal beliefs and biases.
3. Whether the uncertainty associated with resolving ground truth impacts the quantitative accuracy measures.

Each of these is examined below.

4.2 Comparison to other studies

The inferred probability method asks readers to convert verbal certainty statements into precise quantitative probabilities and then uses subjective assessments to resolve ill-defined forecast event statements. One way to examine whether this procedure yields useful accuracy measures is to compare the results obtained above with other studies on similar topics where analysts were asked to assign quantitative probabilities to well-defined events. Although this is not a direct forecast-to-forecast comparison, the overall accuracy profiles should be similar. Below we offer two such comparisons.

Tetlock (2005) summarizes the results from numerous studies with political analysts, "... collapsing over ten thousand predictions for fifty-seven countries across fourteen years." The forecast questions address a variety political analysis issues including numerous political and national stability issues for various regions worldwide. Consequently, the topics covered in Tetlock's forecast studies are similar to the topics covered in the 14 documents examined in this study.

Calibration: Tetlock measures calibration using a statistic called the Calibration Index (CI).[10] For expert political analysts Tetlock found CI=.025. This corresponds roughly to an absolute difference of 0.15 between observed relative frequency and perfect calibration. In our study, where we measured the accuracy of documents written by expert political analysts we found slightly better results. CI=.018 and we directly estimated the absolute difference to be 0.11.

Discrimination: Discrimination refers to the strength of forecast probabilities, where strong forecasts are close to 1.0 or 0.0. Tetlock measures discrimination using a statistic called the Discrimination Index (DI).[11] For expert political analysts Tetlock found DI=.024. For our study we found DI=.066. However, DI can vary greatly as a function of the base rate of forecast events. A "normalized" version of DI (NDI)[12] takes this into account. Tetlock found NDI was approximately 0.20.[13] In our examination of forecast documents written by expert political analysts NDI = 0.28.

Our second comparison is with the results described Mandel et al. (2009). Mandel's study examined the accuracy of a collection of 580 quantitative probabilistic forecasts provided by a group of expert political analysts (intelligence analysts) examining Middle East and African affairs during the period of March 2005 through October 2006. The 14 documents examined in this study addressed the same geographic regions and to some extent the same time periods as Mandel's study. Also the analysts in Mandel's study, like the analysts that authored the documents we examined, were intelligence analysts.

Calibration: In Mandel's study CI=.014, which is slightly better than our finding of CI=.018.

Discrimination: Mandel uses a statistic for measuring discrimination called the Adjusted Normalized Discrimination Index (ANDI), where ANDI is a slight adjustment to the NDI measure.[14] In Mandel's study ANDI=0.58, which is far better than in our study where ANDI = 0.24.

Overall, for both calibration and discrimination, we found that the analysis of the forecast accuracy of documents written by expert political analysts yielded results that were between the results of the two comparative studies examining the accuracy of direct quantitative probability forecasts provided by expert political analysts.

We note here that in this comparison calibration is far more meaningful than discrimination. Recall from the above discussion that these documents exhibit two reporting biases: a tendency to report probable (rather than improbable) events and a tendency to report on difficult to forecast events. In fact 61% of the forecasts (114 out of 187) are in the range 0.7 to 0.9. This narrow range substantially reduces any discrimination measure. Without knowing how the forecast questions were selected in the Tetlock studies or the reporting biases in the Mandel study it's difficult to meaningfully compare discrimination. By contrast, calibration seems less sensitive to reporting biases and therefore provides a more informative comparison. We therefore find it particularly encouraging that the calibration results for all three studies were close.

[10] $CI = (1/N)\cdot(\sum_i N_i\cdot (f_i - d_i)^2)$, where N is the number of observations, N_i is the number of observations at each probability level, f_i is the observed proportion and d_i is the expected proportion.

[11] $DI = (1/N)\cdot(\sum_i N_i\cdot (d_i - d^*)^2)$, where N is the number of observations, N_i is the number of observations at each probability level, d_i is the expected proportion and d^* is the overall proportion of event that occurred.

[12] $NDI = DI/(d^*\cdot (1\text{-}d^*))$

[13] This was inferred from Tetlock's statement "the best human forecasters were hard pressed to predict more that 20% of the total variability in outcomes (using the DI/VI "omniscience" index in the Technical Appendix) ..." The equation in the Technical Appendix is the NDI equation.

[14] $ANDI = (N\cdot NDI - J - 1)/(N - J + 1)$ where J is the number of probability levels.

4.3 Possible impact of reader bias

Eleven of the documents examined here were retrospective studies where readers inferred probabilities years *after* the forecast period had expired. As a practical matter this is how most studies examining the accuracy of forecast documents are likely to be done. However, whether or not its practical, retrospective studies are subject to the criticism that reader inferred probabilities may be heavily influenced by readers' knowledge of whether or not the forecast event occurred. If a reader knew that an event had occurred, then she might be unconsciously inclined to assign a higher inferred probability to that event than if she knew that the event had not occurred. If this is true then events that occurred would receive a higher than warranted inferred probability and events that did not occur would receive a lower than warranted inferred probability—artificially inflating measured accuracy.

To test for this possibility we examined separately the retrospective and prospective studies. In prospective studies the readers inferred probabilities at the beginning of the forecast periods. Since the probabilities are inferred at the beginning forecast period, prospective studies are not subject to the criticism that reader inferred probabilities were biased by their knowledge of whether or not an event had occurred. Consequently, if readers are in fact biased, then they will assign stronger probabilities in retrospective than in prospective studies and will exhibit better accuracy.

Table 3 compares the accuracy statistics for the three prospective studies and the retrospective studies. The first row shows the results for all 11 documents that were the subject of retrospective studies. The second row is for the four documents that match the source (Stratfor) and topic areas of the three prospective studies. Overall the inferred probabilities for the retrospective and prospective studies were equally calibrated, but the prospective studies showed better differentiation. Consequently, there is no evidence in this data to suggest that the accuracy results in the retrospective studies are artificially inflated by the readers' knowledge of the outcomes. And again it is particularly encouraging that calibration results are so similar.

4.4 Impact of errors in ground truth assignments

In academic studies researchers have the luxury of crafting forecasting questions where outcomes can be unambiguously determined at the end of the forecast period. In real world practice, few forecasts meet this criterion. Consequently in our study we used two (and sometimes three) independent raters, who did not see the original forecast document, to judge whether or not the forecast event occurred. Sometimes raters could not judge whether an event occurred and at other times raters would disagree. Overall, when two raters both judged whether or not an event occurred inter rater agreement was only 79%. In the above analysis we used a simple majority rule to resolve ground truth, so if only one rater said an event occurred and the others said "don't know", we determined that the event had occurred. Given a 21% level of disagreement and our willingness to accept the judgment of just one rater, it is reasonable to ask whether and by how much our accuracy statistics are affected by errors in ground truth assignments. As the analysis below shows, our answer is "surprisingly little." Explaining this will take several steps.

First, the reader should appreciate that accounting for errors in ground truth assignments may well *improve* estimated accuracy. To understand this, imagine a set of forecasts where 75% of the forecasted events occurred but the procedure for assigning ground truth is 90% accurate. For this set of forecasted events the expected observed proportion is 70%.[15] Although the true proportion is 75%, the 10% error in assigning ground truth should cause the observed proportion to be lower. Reversing this, if the observed proportion is 70% then we could estimate the true proportion to be 75%.[16]

In general, if the probability of ground truth error is the same for all ground truth judgments, then adjusting for the probability of ground truth error will result in an adjusted proportion that is higher when the observed proportion is above 50% and will result in an adjusted proportion that is lower when the observed proportion is below 50%. Since most calibration curves show underestimates at high probability levels, and overestimates at low probability levels, any adjustment for ground truth error should result in a better calibration score.

Below we describe our procedure for estimating ground truth probabilities and adjusting the calibration curve based on those estimates. We use the data in Table 1 to illustrate the steps.

1. *Calculate inter rater agreement (IRA).* In Table 1 the raters agreed in two of the three cases where they both assigned a Yes or No answer, so IRA=0.667.
2. *Estimate rater accuracy.* Treat each rater as equally accurate and then estimate rater accuracy from IRA. In this case, if each rater is 78.9% accurate in their ground truth judgments, then expected IRA = 66.7%.

[15] Let P_a be the probability that each ground truth assignment is accurate, P_t be the true proportion of events that are true, P_o be the observed proportion of events that are assigned "True" and $E(P_o)$ be the expected value of P_o. Then $E(P_o) = P_t \cdot P_a + (1-P_t) \cdot (1-P_a)$. This is the probability that the event occurred and was correctly assigned "True" plus the probability that the event did not occur and that the event was incorrectly assigned "True".

[16] Specifically $E(P_t) = (1-P_o-P_a)/(1-2 \cdot P_a)$.

Table 3: Comparison of retrospective and prospective studies.

Source of inferred probabilities	Number of forecasts	Difference between mean inferred probability events that occurred and event that didn't occur	Mean absolute error	Mean absolute deviation from perfect calibration
All Retrospective studies	135	0.15	0.37	0.12
Retrospective studies—Stratfor	59	0.15	0.38	0.15
Prospective studies—Stratfor	52	0.37	0.31	0.13

Table 4: Example of adjusting calibration proportion.

	Ground truth probability	Number of cases	Observed proportion	Adjusted proportion
1 Rater	0.789	3	67%	78.7%
2 Raters agree	0.933	2	100%	100%
		Weighted Average =		87.2%

3. *Estimate Ground Truth probability.* Treat each rater as an independent measure of ground truth with the error rate calculated in step 2; then apply Bayes rule to estimate ground truth probabilities. In this case the derived ground truth probability for the first three cases where only one rater answered is 0.789; and where two raters agreed is 0.933.[17]
4. *Estimate ground truth frequency for each calibration level.* This is done by adjusting the observed proportions at each level support (1 rater only, 2 raters agree, etc.), and then taking a weighted average of the adjusted proportions. In Table 1, there were 5 forecasts with an average inferred probability around 0.85; where four of those events occurred. So the observed proportion was 80%. But as shown in Table 4, the adjusted proportion is 87.2%.

Across the 14 documents in this study, inter rater agreement was 79%, from which we deduce an estimated accuracy for each ground truth rater of 88%. So for each inferred probability level we used the procedure illustrated in Table 4 to derive an adjusted proportion. For example, there were 51 forecast events for which the inferred probability was 0.9; where 41 of the 51 (80.4%) were rated as True. Applying the procedure illustrated in Table 4 to those 51 events yielded an adjusted proportion of 84.2%.

Figure 2 shows the calibration curve for both the observed and adjusted proportions for all 187 forecasts. As can be seen there is very little difference and all of the above mentioned metrics yield nearly identical results.

The procedure we use to estimate ground truth probabilities makes several assumptions; all raters are equally accurate, ratings of both occurrence and non-occurrence of an event are equally accurate and equal prior probabilities.[18] Here we do not argue the merits of these assumptions, but rather simply note that it is straightforward to adjust for possible errors in ground truth ratings. And that, at least for our data set, this adjustment has little impact on estimated accuracy.

Discussion

This paper presents a method, called the inferred probability method, for quantitatively measuring the accuracy of forecasts in documents that use imprecise language to describe both forecast events and forecast certainties. Because many real world forecasts are expressed with verbal imprecision we believe that the use of this method will substantially expand the range of forecasts and forecasting methods that are amenable to empirical analysis.

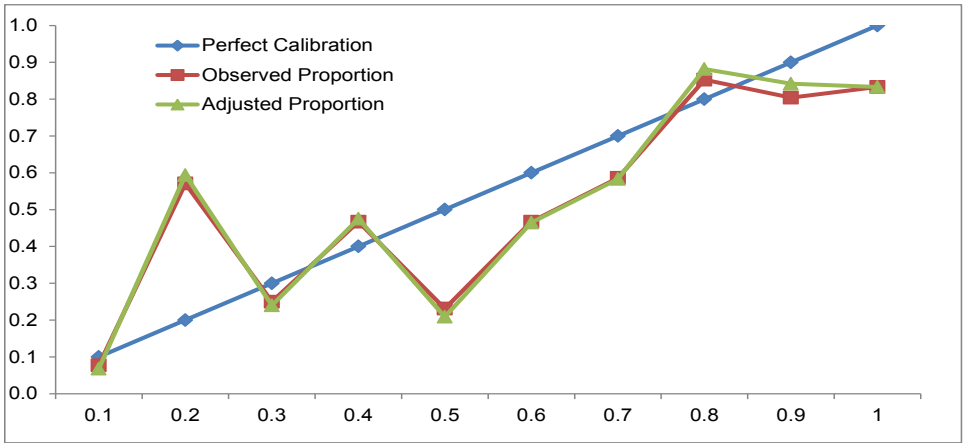
In an effort to test its applicability, we applied the inferred probability method to 14 documents that examine significant and complex political events, including two declassified National Intelligence Estimates, which are considered the premier analysis product of the United States Intelligence Community. Our test focused on three criteria:

1. Whether the inferred probability method yielded accuracy results that are in the same range as more tra-

[17] We used Bayes rule with a prior of 0.5 where each rater is treated as a conditionally independent measure of ground truth.

[18] In our view these same assumptions that are implicitly made by many studies that aggregate the ratings of multiple raters.

Figure 2: Observed and adjusted calibration curves.



ditional forecasting studies in the same general topic area.

2. Whether the accuracy results were biased by a readers' knowledge of the topic area, and
3. Whether the accuracy results were sensitive to errors in assigning ground truth.

When applied to 14 documents forecasting complex international political affairs we found that:

1. Calibration results were similar to those found in studies where experts directly provided quantitative probabilities for clearly worded forecast events.
2. The accuracy results were largely the same whether the studies were retrospective (probabilities inferred after the end of the forecast period) or prospective (probabilities inferred at beginning of the forecast period); indeed the prospective studies yielded slightly better results.
3. A statistical analysis of the impact of possible errors in ground truth assignments suggests that such errors have little impact on measured accuracy.

Overall, we believe these results support a claim that the inferred probability method can be used to routinely evaluate the accuracy of forecast documents where forecast events and certainties were expressed with imprecision. Since many significant forecasts are expressed with verbal imprecision, we believe that routine use of the inferred probability method could substantially expand the evidence-base and relevance of forecasting research.

Although the inferred probability method appears to have considerable utility, different researchers may choose to instantiate this method differently. Our particular instantiation reflects two key choices that we believed appropriate to our research objectives and substantive domain of interest, but would vary for other applications.

Our first choice was to include all forecast events in a document and to write the event statement in the same language as was originally expressed in the document. We did this even when the forecast event statement in the document was egregiously vague. We left it to the ground truth raters to tell us if the event statement was too vague to resolve. We choose this route because our objective was to evaluate the overall accuracy of these documents and so we did not want to arbitrarily exclude portions of the document. Other researchers may have different objectives which may lead them to use well-defined forecast events. For example, researchers may want to directly compare different sources of forecasts, such as different forecast documents, on a common set of forecasting questions. For such studies readers can infer probabilities for forecast events even though the event statement is not expressed in exactly the same words that are found in each document. In such comparative studies there would be no reason to use anything other than well-defined forecast event statements. The use of well-defined forecast events would also remove concerns about errors in ground truth assignments. Furthermore, if all of the documents or other sources of forecasts are measured against the same forecast questions, then discrimination and error score measures can be meaningfully applied.

Our second choice was to use readers who were experienced professionals with some substantive knowledge of the forecast topic areas. We choose these readers because we felt that they reflected the population of serious readers of these documents; and we were particularly interested in the accuracy of interpretations of such readers. However, their substantive knowledge also increased the potential for biased inferred probabilities, where they might assign a higher inferred probability to events that they knew had occurred. Although our comparison of ret-

rospective and prospective studies suggests that this was not an issue, we do not claim that this is a general result. In future studies it would be wise to use a mixture of readers, some of whom should be uninformed on the subject matter of the forecast document. Then the impact of substantive knowledge can be measured.

No matter how the inferred probability method is used or modified, we believe that this general approach can substantially expand the range of forecasts that are subject to rigorous empirical assessment.

References

Alvarez, A., Cohn, L. D., & Vazquez, M. E. C. (2009). Quantifying risk: Verbal probability expressions in Spanish and English. *American Journal of Health Behavior*, *33*, 244–255.

Africa: Attracting outside interest, But conflicts continue. (2007, Jan 18). In *2007 annual forecast: Time to look inward—Part II*. Retrieved December 9, 2009 from Stratfor website: http://www.stratfor.com/forecast/2007-annual-forecast-time-look-inward-part-ii.

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

Beyth-Marom, R. (1983). How probable is probable? Numerical translation of verbal probability expressions. *Journal of Forecasting*, *1*, 257–269.

Further JID Forecasts. (2006, Jan 11). In *Jane's Intelligence Digest*. Retrieved December 9, 2009 from Jane's web site: https://globalsso.ihs.com/KeystoneSTS/SSOLogin/Login.aspx?altertheme=janes&theme=ENERGY&ReturnUrl=https%3a%2f%2fglobalsso.ihs.com%2fKeystoneSTS%2fSaml2%2fDefault.aspx%3faltertheme%3djanes.

Gardner, D. (2010). *Future Babble: Why Expert Predictions Fail – and Why We Believe Them Anyway*. Toronto: McClelland and Stewart.

Kennedy, D. T., & Tavana, M. (1997). An applied study using the Analytic Hierarchy Process to translate common verbal phrases to numerical probabilities. *Journal of Behavioral Decision Making*, *10*, 133–150.

Larijani's Departure Fuels Iran Power Struggle. (2007, Nov 8). Retrieved December 9, 2009, from Jane's web site: https://globalsso.ihs.com/KeystoneSTS/SSOLogin/Login.aspx?altertheme=janes&theme=ENERGY&ReturnUrl=https%3a%2f%2fglobalsso.ihs.com%2fKeystoneSTS%2fSaml2%2fDefault.aspx%3faltertheme%3djanes.

Mandel, D. R., Barnes, A., & Hannigan, J. (2009, February). A calibration study of an intelligence assessment division. Paper presented at the *Global futures forum community of interest for the practice and organization of intelligence Ottawa—What can the cognitive and behavioural sciences contribute to intelligence analysis? Towards a collaborative agenda for the future.* Meech Lake, Quebec.

Middle East and South Asia: Accommodation. (2006, Jan 17). In *Annual forecast 2006: The year of great and near-great-powers—Part I*. Retrieved December 9, 2009 from Stratfor Website: http://www.stratfor.com/forecast/annual-forecast-2006-year-great-and-near-great-powers-part-i.

Middle East: Pivoting on Developments Between Washington and Tehran (2007, Jan 18). *In 2007 annual forecast: Time to look inward—Part I*. Retrieved December 9, 2009 from Stratfor website: http://www.stratfor.com/forecast/2007-annual-forecast-time-look-inward-part-i.

Middle East/South Asia. (2011, Jan 13). In *Annual forecast 2011*. Retrieved January 17, 2011 from Stratfor website: http://www.stratfor.com/forecast/annual-forecast-2011.

Prospects for Iraq's stability: A challenging road ahead. (2007, Jan). Retrieved March 4, 2010 from Director of National Intelligence website: http://dni.gov/files/documents/Newsroom/Reports%20and%20Pubs/20070202_release.pdf.

Sub-Saharan Africa: Shifting of powers. (2006, Jan 17). In *Annual forecast 2006: The year of great and near-great-powers—Part II*. Retrieved December 9, 2009 from Stratfor Website: http://www.stratfor.com/forecast/annual-forecast-2006-year-great-and-near-great-powers-part-ii.

Sub-Saharan Africa. (2011, Jan 13). In *Annual forecast 2011*. Retrieved January 17, 2011 from Stratfor website: http://www.stratfor.com/forecast/annual-forecast-2011.

Tetlock, P. (2005). *Expert political judgment*. Princeton: Princeton University Press.

Trends in global terrorism: Implications for the US. (2006, April). Retrieved March 4, 2010 from Director of National Intelligence website: http://dni.gov/files/documents/Newsroom/Press%20Releases/2006%20Press%20Releases/Declassified_NIE_Key_Judgments.pdf.

US and Iran: Road map to conflict. (2007, Feb 16). Retrieved December 9, 2009 from Jane's web site: https://globalsso.ihs.com/KeystoneSTS/SSOLogin/Login.aspx?altertheme=janes&theme=ENERGY&ReturnUrl=https%3a%2f%2fglobalsso.ihs.com%2fKeystoneSTS%2fSaml2%2fDefault.aspx%3faltertheme%3djanes.

Wallsten, T. S., Budescu, D. V., Erev, I., & Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making*,

10, 243–268.

Wallsten, T., Budescu, D., & Zwick, R. (1992). Comparing calibration and coherence of numerical and verbal probability judgments. *Management Science*, *39*, 176–190.

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