DRDC-RDDC-2014-P72
September 2014

Classification: Major: Biological Sciences; Minor: Psychological and Cognitive Sciences

Accuracy of Forecasts in Strategic Intelligence

Short title: Accuracy of Forecasts in Strategic Intelligence

David R. Mandel[a1], Alan Barnes[b]

[a]Socio-Cognitive Systems Section, DRDC, Toronto Research Centre, 1133 Sheppard
Avenue West, Toronto, Ontario M3K 2C9 Canada

[b]Formerly of the Intelligence Assessment Secretariat, Privy Council Office

[1]To whom correspondence should be addressed. E-mail: drmandel66@gmail.com

Keywords: forecasting, accuracy, strategic intelligence

Abstract

The accuracy of 1,514 strategic intelligence forecasts abstracted from intelligence reports was assessed. The results show that both discrimination and calibration of forecasts was very good. Discrimination was better for senior (versus junior) analysts and for easier (versus harder) forecasts. Miscalibration was mainly due to underconfidence such that analysts assigned more uncertainty than needed given their high level of discrimination. Underconfidence was more pronounced for harder (versus easier) forecasts and for forecasts deemed more (versus less) important for policy decision-making. In spite of the observed underconfidence, there was a paucity of forecasts in the least informative 0.4-0.6 probability range. Recalibrating the forecasts substantially reduced underconfidence. The findings offer cause for tempered optimism about the accuracy of strategic intelligence forecasts and indicate that intelligence producers aim to promote informativeness while avoiding overstatement.

Significance Statement

Forecasting is an important part of strategic intelligence, offering policy makers indications about probable future conditions and aiding sound decision-making. Nevertheless, there has not been a concerted effort to study the accuracy of intelligence forecasts over a large set of assessments. This study applied quantitative measures of forecasting accuracy commonly used in other areas of expert forecasting to over 1,500 strategic intelligence forecasts spanning approximately 6 y of intelligence production by a strategic assessment unit. The findings revealed a high level of discrimination and calibration skill. We also show how calibration may be improved through post-forecast transformations. The study illustrates the utility of proactively applying scoring rules to intelligence forecasts as a means of outcome-based quality control.

\body

Strategic intelligence assists high-level decision-makers, such as senior government leaders, in understanding the geopolitical factors shaping the world around them. Such intelligence can help decision-makers anticipate future events, avoid strategic surprises, and make informed decisions. Policy-neutral intelligence that is timely, relevant, and accurate can be of significant value to decision-makers (1). Although not all intelligence is predictive, forecasts are an important part of intelligence, serving to reduce uncertainty about future events for decision-makers (2). Forecasts comprise a substantial part of the type of judgment that Sherman Kent (widely regarded as the father of modern intelligence analysis) identified as most informative—namely, judgments that go beyond the information given (3). A question arising, then, is how good are intelligence analysts at forecasting geopolitical events?

The answer to that question should be of value to multiple stakeholders. First, intelligence consumers should want to know how good the forecasts they receive actually are. This should guide the weight they assign to them and the trust they place in their sources. Second, intelligence directors directly accountable for analytic quality control should want to know how well their analysts are doing. An objective scorecard might mitigate "accountability ping-pong" pressures in which the intelligence community reactively shifts its tolerance levels for false-positive and false-negative errors in order to "now get it right" (4). Third, analysts, an intellectually curious breed, should want to know how good their forecasts are. Beyond satisfying curiosity, receiving objective performance feedback on a regular basis can encourage adaptive learning by revealing judgment characteristics (e.g., overconfidence) that would be hard to detect from case-specific reviews (5). Finally, citizens should want to know how well the intelligence community is doing. Not only does personal security depend on national security, intelligence is a costly enterprise, running into the tens of billions each year in the US (6).

Despite good reasons to proactively track the accuracy of intelligence forecasts, intelligence organizations seldom keep an objective scorecard of forecasting accuracy (7). There are many reasons why they do not do so. First, analysts seldom use numeric probabilities, which lend themselves to quantitative analyses of accuracy. Many analysts, including Kent's "poets" (3), are resistant to the prospect of doing so (8). Second, intelligence organizations do not routinely track the outcomes of forecasted events, which are needed for objective scorekeeping. Third, only recently have behavioral scientists offered clear guidance to the intelligence community on how to measure the quality of human judgment in intelligence analysis (9-12). Finally, there may be some apprehension within the community regarding what a scorecard might reveal.

Such apprehension is understandable. The intelligence community has been accused of failing to "connect the dots" (13), which would seem to be a requisite skill for good forecasting. Nor does literature on human forecasting accuracy inspire high priors for success. It has long been known that people are miscalibrated in their judgment (14), tending towards overconfidence (15). Tetlock's study of political experts' forecasts showed that, although experts outperformed undergraduates, even the best human forecasters—the Berlinian foxes—were left in the dust when compared to the savvier statistical models he tested (16), recalling earlier studies showing how statistical models outperform human experts' predictions (17-18).

Nevertheless, it would be premature to draw a pessimistic conclusion about the accuracy of intelligence forecasting without first-hand examination of a sufficient number of actual intelligence forecasts. As well, some experts, such as meteorologists (19-20) and bridge players (21), are very well calibrated. Intelligence analysts share some similarities with these experts. For instance, their forecasts are a core product of their expertise. This is not true of all experts, not even those who make expert probabilistic judgments. For instance, physicians may be overconfident (22), but, if their patients recover, they are unlikely to question their forecasts. Likewise for lawyers and their clients (23). In these cases, the success of experts' post-forecast interventions will matter most to clients. However, analysts do not make policy decisions. Their expertise is shown directly by the quality of their policy-neutral judgments.

Here, we report the findings of a long-term field study of the quality of strategic intelligence forecasts. We examined an extensive range of intelligence reports produced by a strategic intelligence unit over an approximate 6-y period (March 2005 to December 2011). From each report, every judgment was coded for whether it was a forecast given that not all judgments are predictive (e.g., some are explanatory). Outcomes of the forecasted events were tracked, enabling us to quantify several aspects of forecast quality. Those aspects are roughly grouped into two types of indices. The first, discrimination [also called resolution (24)], refers to how well a forecaster separates occurrences from non-occurrences. The second, calibration [or reliability (24)], refers to the degree of correspondence between a forecaster's subjective probabilities and the observed relative frequency of event occurrences. We use multiple methods for assessing discrimination and calibration, including receiver operating characteristic (ROC) curve analysis (25) Brier score decomposition (24, 26-27), and binary logistic models for plotting calibration (28-29), thus allowing for a wide range of comparison to other studies.

We examined discrimination and calibration for the overall set of forecasts, as well as whether forecast quality was influenced by putative moderating factors, including analyst experience, forecast difficulty, forecast importance, and temporal window size. Would senior analysts forecast better than junior analysts? The literature offers mixed indications. For example, while expertise benefitted the forecasts of bridge players (21), it had no effect on expert political forecasters (16). Would easier forecasts be better than harder forecasts? Calibration is often sensitive to task difficulty such that harder problems yield overconfidence that is attenuated for easier problems—what is commonly termed the hard-easy effect (30). Would forecast quality be better for the more important forecasts? This question is of practical significance because, while intelligence organizations strive for accuracy in all their assessments, they are especially concerned about "getting it right" on questions of greatest importance to their clients. Finally, would accuracy depend on whether the temporal window size was smaller (0-6 months) or larger (6 months to about 1 y)? Larger windows might offer better odds of the predicted event occurring, much as a larger dartboard would be easier than a smaller one to hit. In addition to our descriptive aims, we were interested in the possible prescriptive value of the work. Prior research has shown that forecasting quality may be improved through post-forecast transformations. Some approaches require the collection of additional forecasts. For example, using small sets of related probability estimates, Karvetski, Olson, Mandel, and Twardy (31) were able to improve discrimination and calibration skill by transforming judgments so that they minimized internal incoherence. Other studies have shown that calibration can be improved through transformations that do not require collecting additional judgments (32-33). To the extent that forecasts in this study were miscalibrated, we planned to explore the potential benefit and usability of transformations that do not require the collection of additional forecasts.

Results

Discrimination. Fig. 1 shows a discrimination diagram of the forecast-outcome cross-tabulation data. The data are arrayed such that the four quadrants of the diagram show the frequency of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) when forecast probabilities ≤ 0.5 indicate event non-occurrence and probabilities > 0.5 indicate event occurrence. Fig. 1 shows that most forecasts (94%) were correctly classified as either TP or TN cases. As well, few cases were in the middle (0.4-0.6) range of the probability scale. This is advantageous because such high-uncertainty forecasts are less informative for decision-makers.

Fig. 2 shows the ROC curve, plotting the TP rate (sensitivity) as a function of the FP rate (1 – specificity). A useful measure of discrimination skill is the area under the ROC curve, A (25), the proportion of the total area of the unit square defined by the two axes of the ROC curve. A can range from 0.5 (i.e., the area covered by the 45° no-discrimination line) to 1.0 (perfect discrimination). A was 0.940 (standard error, $SE = 0.007$), which is very good.

To assess the effect of the putative moderators, A_i was compared across levels of each moderator using the following Z-score formula (34):

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

Discrimination was significantly higher for senior analysts ($A_1 = 0.951$, $SE = 0.008$) than for junior analysts ($A_2 = 0.909$, $SE = 0.017$), $Z = 2.24$, $P = 0.025$. Likewise, discrimination was higher for easier forecasts ($A_1 = 0.966$, $SE = 0.010$) than for harder forecasts ($A_2 = 0.913$, $SE = 0.011$), $Z = 3.57$, $P < 0.001$. Discrimination did not differ by importance or temporal window size ($P > .25$).

Finally, discrimination was also very good when the forecast probabilities were taken into account using Brier score decomposition. The Brier score, B, is the squared deviation between a forecast and the outcome coded 0 (non-occurrence) or 1 (occurrence). The mean Brier score is a proper scoring rule for probabilistic forecasts:

$$\bar{B} = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2, \qquad (2)$$

where N is the total number of forecasts (1,514 in this study), f_i is the subjective probability assigned to the ith forecast, and o_i is the outcome of the ith event, coded 0 or 1. However, because $\bar{B}$ is affected not only by forecaster skill, but also by the uncertainty of the forecasting environment, it is usually decomposed into indices of variance (VI), discrimination (DI), and calibration (CI):

$$\bar{B} = VI - DI + CI, \text{ where} \tag{3}$$

$$VI = \bar{o}(1-\bar{o}), \tag{4}$$

$$DI = \frac{1}{N}\sum_{k=1}^{K} N_k(\bar{o}_k - \bar{o})^2, \text{ and} \tag{5}$$

$$CI = \frac{1}{N}\sum_{k=1}^{K} N_k(f_k - \bar{o}_k)^2. \tag{6}$$

In Eqn. 4-6, $\bar{o}$ is the base rate or overall relative frequency of event occurrence, K is the number of forecast categories (9 in this study), N_k is the number of forecasts in category k, f_k is the subjective probability assigned to forecasts in category k, and $\bar{o}_k$ is the relative frequency of event occurrence in category k. In this study, $\bar{B}$= 0.074, VI = 0.240, and DI = 0.182. Because DI is upper-bounded by VI, it is common to normalize discrimination:

$$\eta^2 = DI / VI. \tag{7}$$

An adjustment is sometimes made to this measure (26). However, it had no effect in this study. Normalized discrimination was very good, with forecasts explaining 76% of outcome variance, $\eta^2 = 0.758$.

Calibration. The calibration index, CI, which sums the squared deviations of forecasts and the relative frequencies of event occurrence for each forecast category, is perfect when $CI = 0$. In this study, $CI = 0.016$. An alternative measure of calibration called calibration-in-the-large (27), CI_L, is the squared deviation between the mean forecast and the base rate of event occurrence over all categories:

$$CI_L = \bar{f} - \bar{o}. \tag{8}$$

Calibration in the large was virtually nil, $CI_L = 3.60\text{e}{-5}$. Thus, both indices showed that forecasts were well calibrated.

Calibration curves were modeled using a generalized linear model with a binary logistic link function. Event outcomes were first modeled on forecast, difficulty, importance, experience, temporal window size, and all two-way interactions with forecast. Two predictors were significant: forecast and the forecast × difficulty interaction. Using forecast and difficulty as predictors, the model fit for forecast only and forecast with the interaction term were compared using Akaike Information Criterion [AIC (35)]. The model including the interaction term improved model fit (AIC = 118.14) over the simpler model (AIC = 148.47). Fig. 3 shows the calibration curves plotting mean predicted probability of the event occurring as a function of forecast and difficulty. It is evident that in both the easier and harder sets of forecast, miscalibration was mainly due to underconfidence, as revealed by the characteristic S-shape of the curves. This pattern of calibration is also referred to as underextremity bias because the forecasts are not as extreme as they should be (36). In the easier set, underconfidence is more pronounced for forecasts above 0.5, whereas underconfidence is more pronounced for forecasts below 0.5 in the harder set.

A signed calibration of confidence index, *CCI*, was computed as follows, such that negative values indicated underconfidence and positive values indicated overconfidence:

$$CCI = \frac{1}{N}\left(\sum_{k=1}^{K} N_k(\bar{o}_k - f_k) \text{ iff } f_k < 0.5 + \sum_{k=1}^{K} N_k(f_k - \bar{o}_k) \text{ iff } f_k > 0.5\right). \tag{9}$$

This measure omits the small number of cases where $f_k = 0.5$. In this study, $CCI = –0.076$ ($s = 0.096$), therefore indicating underconfidence. The 99% confidence interval on the estimate ranged from –0.082 to –0.070, placing zero well outside its narrow range. Given the non-normality of subsample distributions of underconfidence, we examined the effect of the moderators on underconfidence using Mann-Whitney *U* tests. There was no effect of experience or temporal window size ($P > 0.15$). However, underconfidence was greater for harder forecasts (median = –0.106) than for easier forecasts (median = –0.044), $Z = –13.39$, $p < .001$. Underconfidence was also greater for forecasts of greater importance (median = –0.071) rather than lesser importance (median = –0.044), $Z = –4.49$, $P < 0.001$. Given that difficulty and importance were positively related ($r = .17$, $P = 0.001$), we tested whether each had unique predictive effects on underconfidence using ordinal regression. Both difficulty ($b = 1.26$, $SE = 0.10$, $P < 0.001$) and importance ($b = 0.32$, $SE = 0.12$, $P = .007$) predicted underconfidence, together accounting for roughly an eighth of the variance in underconfidence (Nagelkerke pseudo-$R^2 = 0.124$).

Sensitivity. The high degree of forecasting quality raises questions about the sensitivity of the results to excluded cases. There were two types of exclusions we explored: cases with ambiguous outcomes and cases without numeric probabilities. The first type accounted for about 20% of numeric forecasts and typically occurred when the event in question was not defined crisply enough to make an unambiguous determination of outcome. In such cases, the coders either assigned partial scores indicating evidence in favor of occurrence or non-occurrence, or else an unknown code. We dummy coded partial non-occurrences as 0.25, unknown cases as 0.50, and partial occurrences as 0.75 and recomputed the mean Brier score. The resulting value was 0.090. A comparable analysis was undertaken to assess the effect of excluding the roughly 16% of forecasts that did not have numeric probabilities assigned. In these cases, terms such as could or might were used and these were deemed to be too imprecise to assign numeric equivalents. For the present purpose, we assume that such terms are likely to be interpreted as being a fifty-fifty call (37) and, accordingly, we assign a probability of 0.50 to those cases. The recomputed mean Brier score was 0.097. Given that both recomputed Brier scores are still very good, we can be confident that the results are not due to case selection biases.

Recalibration. Given that forecasts exhibited underextremity bias, we applied a transformation that made them more extreme. Following (32-33), which draws on Karmarker's earlier formulation (38), we used the following transformation:

$$t = \frac{f^a}{f^a + (1-f)^a}, \tag{10}$$

where t is the transformed forecast and a is a tuning parameter. When $a > 1$, the transformed probabilities are made more extreme. We varied a by 0.1 increments from 1-3 and found that the optimal value for minimizing *CI* was 2.2. Fig. S1 plots *CI* as a function of a. Fig. S2 shows the transformation function when $a = 2.2$. Recalibrating the forecasts in this manner substantially improved calibration, $CI = 0.0018$. To assess the degree of improvement, it is useful to consider the square root of *CI* in percentage format, which is the mean absolute deviation from perfect calibration. This value is 12.7% for the original forecasts and 4.2% for the transformed forecasts, a 66.9% decrease in mean absolute deviation.

Although the Karmarker transformation improved calibration, the resulting forecast probabilities have no analog on the numeric scale used by the intelligence unit that we studied. To assess whether a more feasible recalibration method that uses the original scale points may be of value, the transformed values were mapped onto their closest original scale point, resulting in a remapping of the original forecasts, f_k, to the new transformed forecasts, t^*_k as follows for $[f_k, t^*_k]$: [0, 0], [0.1, 0], [0.25, 0.1], [0.4, 0.25], [0.5, 0.5], [0.6, 0.75], [0.75, 0.9], [0.9, 1], and [1, 1]. With this feasibility adjustment, $CI = 0.0021$. The mean absolute deviation from perfect calibration was 4.6%, a slight difference from the optimized Karmarker transformation. Fig. 4 shows the calibration curve for the 1,514 forecasts before and after remapping, while Fig. S3 shows the curves before and after the Karmarker transformation was applied. The two figures are almost indistinguishable.

Discussion

Our findings warrant tempered optimism about the quality of strategic intelligence forecasts. The forecasts fared exceptionally well in terms of discrimination and calibration, and these results were not particularly sensitive to case exclusions. The results provide a stark comparison to Tetlock's (16) findings. Whereas the best political forecasters in his sample explained about 20% of the outcome variance, forecasts in this study explained 76% of the variance in geopolitical outcomes. Likewise, the mean absolute deviation from perfect calibration (the square root of the calibration index) was 25% smaller in our study than in Tetlock's, reflecting better calibration. Experience had no effect in his study, but analytic experience led to a practically beneficial improvement in discrimination in this study. Finally, whereas experts in his study showed overconfidence, forecasts in this study were underconfident. Although unbiased confidence is optimal, the cost of overconfidence in intelligence will usually outweigh that of underconfidence. Underconfident forecasting occurs when analysts discriminate better than they think they can. Although underconfidence reduces informational value for decision-makers by expressing forecasts too timidly, it signals good judgment. Overconfidence, in contrast, indicates that forecasts are communicated with unwarranted certainty, implying that they are more error-prone than experts predict.

Several of the findings suggest that intelligence producers adhere to a professional norm of promoting informativeness while avoiding overstatement. First, there were few uninformative forecasts near maximum uncertainty (probabilities of 0.4-0.6). If analysts were only concerned with playing it safe, we would likely have seen a bulge rather than a trough in that region. Second, underconfidence was more pronounced when the issues were relatively complex and more important for policy makers. These variations are consistent with a norm of caution, which both difficulty and consequence of judgment should heighten. Indeed, the effect of difficulty may otherwise appear surprising given that harder problems usually produce greater overconfidence (29). A norm of caution may serve the intelligence community well in the absence of systematic feedback on the forecasting quality. With such feedback, adjustments to cautionary normative pressure could lead to crisper indications as analysts and directors become aware of characteristics such as good discrimination (5). That would, in turn, better serve the aim of reducing uncertainty in the mind of the decision-maker (2).

A deeper account of the present findings might trace normative pressures for cautious, yet informative assessments to analysts' accountability to multiple, skeptical audiences. Accountability pressure has been shown to reduce overconfidence in predicting personality attributes (39), reduced over-attribution bias (40), deeper information processing (41) and better awareness of the informational determinants of one's choices (42). As Arkes and Kajdasz (43) have already noted about a preliminary summary of our findings presented to the National Research Council Committee on Behavioral and Social Science Research to Improve Intelligence Analysis for National Security, the strategic analysts who made the forecasts studied here had much more severe accountability pressures than experts in Tetlock's (16) study and in other academic research. Experts in such studies know that their forecasts will be anonymously analyzed in the aggregate. In contrast, analysts are personally accountable to multiple, skeptical audiences. They have to be able to defend their assessments to their directors. And, together with those directors, they need to be able to "speak truth to power." Analysts are acutely aware of the multiple costs of getting it wrong, especially on the most important issues.

As noted earlier, underconfidence, as a source of miscalibration, is the lesser of two evils. It is also to a large extent correctable. We showed that miscalibration could be attenuated post-forecast in ways that are organizationally usable. Using the Karmarker transformation as a guide, we remapped forecasts to the organizational scale value that had come closest to the transformed value. This procedure yielded a substantial 63.8% decrease in mean absolute deviation from perfect calibration. If intelligence organizations were able to discern stable biases in forecasting, such corrective measures could potentially be implemented between forecast production and finished intelligence dissemination to consumers.

Of course, recalibration, like any model-fitting exercise, requires caution. Aside from the philosophical matter of whether the transformed values still represent subjective probabilities (44), care ought to be taken not to overfit or overgeneralize the approach. This is especially likely in cases where the recalibration rules are based on small sets of forecasts or where attempts are made to generalize to other contexts. Accordingly, we would not advise a different intelligence organization to similarly make their forecasts more extreme without having first studied their own forecast characteristics. A further complication may arise if intelligence consumers have already learned to recalibrate a source's forecasts. For instance, if policy makers know that the forecasts they receive from a given source tend to be underconfident, they may intuitively adjust for that bias by decoding the forecast in more extreme terms (33). Thus, a change in procedure that resulted in debiased forecasts would somehow have to be indicated to consumers so that they do not continue to correct the corrected forecasts. What this work suggests, however, is that, at least in principle, such correctives are doable.

Materials and Methods

Forecast data were acquired from internal versions of intelligence reports produced by the Middle East and Africa Division of the Intelligence Assessment Secretariat (IAS). IAS is the Canadian government's strategic intelligence analysis unit, providing original, policy-neutral assessments on foreign developments and trends for the Privy Council Office and senior clients across government.

The internal versions of reports included numerical probabilities assigned to unconditional forecasts, which could take on values of 0, 1, 2.5, 4, 5, 6, 7.5, 9, or 10 (out of 10). Such numbers were only assigned to judgments that were unconditional forecasts and they were not included in finished reports. Analysts made the initial determination of whether their judgments were unconditional forecasts, conditional forecasts (e.g., of the form "if *X* happens, then *Y* is likely to occur"), or explanatory judgments (e.g., judgments of probable cause), and those were subsequently reviewed and, if necessary, discussed with the director, who had the final say.

From 3,881 judgments, classifications were as follows: 59.7% (2,315) were unconditional forecasts, 15.0% were conditional forecasts, and 25.3% were explanatory judgments. From the unconditional forecasts, 83.8% (1,943) had numeric probabilities assigned. The remaining cases were not assigned numeric probabilities because the verbal expressions of uncertainty used were deemed too imprecise to warrant a numeric translation.

Two subject-matter experts unaffiliated with the division coded outcomes. Each coder handled a distinct set of forecasts, although both coded the same 50 forecasts to assess reliability, which was at 90% agreement. While having unambiguous, mutually exclusive and exhaustive, outcome possibilities for all forecasts is desirable (16), such control was impossible here. In 80.4% (i.e., 1,562/1,943) of cases, forecasts were articulated clearly enough that the outcome could be coded as an unambiguous non-occurrence or occurrence. In the remaining cases, however, the outcome was ambiguous, usually because the relevant forecast lacked sufficient precision and actual conditions included a mix of "pro" and "con" evidence. In such cases, outcomes were assigned to "partial non-occurrence," "partial occurrence," or "unknown" categories, depending on the coders' assessments of the balance of evidence. However, the primary analyses are restricted to the unambiguous set.

Four putative moderating variables were also coded. The second author coded analyst experience, assigning the analyst to either a junior or senior category. In some cases, an analyst who was junior in a given year was reclassified as senior in subsequent years. Thirty cases made in teams were excluded. Coders coded forecast difficulty (easier/harder), forecast importance (lesser/greater), and temporal window size (less than/greater than 6 months). The second author reviewed the codes and discrepant assignments were resolved by discussion. An additional 18 cases were excluded because of missing data on one or more variables. The final set of cases included 1,514 forecasts.

Acknowledgments

This work was supported by DRDC Applied Research Program project 15dm "Understanding and Augmenting Human Analytic Capabilities" and Natural Sciences and Engineering Research Council of Canada Discovery Grant 249537-07. We thank Quan Lam, Philip Omorogbe, and Karen Richards for their research assistance.

References

1. Treverton GF (2008) in *Analyzing Intelligence: Origins, Obstacles, and Innovations*, eds George RZ, Bruce JB (Georgetown University Press, Washington) pp 91-106.

2. Fingar T (2011) in *Intelligence Analysis: Behavioral and Social Scientific Foundations,* eds Fischhoff B, Chauvin C (National Academies Press, Washington) pp 3-27.

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

4. Tetlock PE, Mellers BA (2011) Intelligent management of intelligence agencies: Beyond accountability ping-pong. *Am Psychol* 66: 542-554.

5. Rieber S (2004) Intelligence analysis and judgmental calibration. *International Journal of Intelligence and CounterIntelligence* 17: 97–112.

6. Reuters (October 28, 2011) *Fiscal 2011 U.S. intelligence budget was 54.6 bln.* Available at: http://www.reuters.com/article/2011/10/28/usa-intelligence-budget-idUSN1E79R1CN20111028

7. Betts RK (2007) *Enemies of Intelligence: Knowledge and Power in American National Security* (Columbia University Press, New York), 241 pp.

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

9. Fischhoff B, Chauvin C, eds (2011) *Intelligence Analysis: Behavioral and Social ntelligence analysis: Behavioral and social scientific foundations* National Academies Press, Washington), 338 pp.

10. Derbentseva N, McLellan L, Mandel DR (2011) *Issues in Intelligence Production: Summary of Interviews with Canadian Intelligence Managers.* DRDC Toronto Technical Report 2010-144.

11. National Research Council (2010) *Field Evaluation in the Intelligence and Counterintelligence Context: Workshop Summary* (National Academies Press, Washington), 114 pp.

12. National Research Council (2011) *Intelligence Analysis for Tomorrow: Advances from the Behavioral and Social Sciences* (National Academies Press, Washington) 102 pp.

13. National Commission on Terrorist Attacks upon the United States (2004) *The 9/11 Commission Report: Final report of the National Commission on Terrorist Attacks upon the United States* (US Government Printing Office, Washington), 585 pp.

14. Lichtenstein S, Fischhoff B, Phillips LD (1982) in *Judgment under Uncertainty: Heuristics and Biases,* eds Kahneman D, Slovic P, Tversky A (Cambridge University Press, New York), pp 306-334.

15. Fischhoff B, Slovic P, Lichtenstein S (1977) Knowing with certainty: The appropriateness of extreme confidence. *J Exp Psychol Hum Percept Perform* 3: 552-564.

16. Tetlock PE (2005) *Expert Political Judgment: How Good is It? How Can We Know?* (Princeton University Press), 321 pp.

17. Dawes RM (1979) The robust beauty of improper linear models in decision making. *Am Psychol* 34: 571-582.

18. Meehl PE (1954) *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence* (University of Minnesota Press), 149 pp.

19. Murphy AH, Winkler RL (1984) Probability forecasting in meteorology. *J Am Stat Assoc* 79: 489-500.

20. Charba JP, Klein WH (1980) Skill in precipitation forecasting in the National Weather Service. *Bull Amer Meteor Soc* 61: 1546-1555.

21. Keren G (1987) Facing uncertainty in the game of bridge: A calibration study. *Organ Behav Hum Decis Process* 39: 98-114.

22. Berner ES, Graber ML (2008) Overconfidence as a cause of diagnostic error in medicine. *Am J Med* 121: S2-S23.

23. Goodman-Delahunty J, Granhag PA, Hartwig M, Loftus EF (2010) Insightful or wishful: Lawyers' ability to predict case outcomes. *Psychol Public Policy Law* 16: 133-157.

24. Murphy AH (1973) A new vector partition of the probability score. J Appl Meteorology 12(4): 595–600.

25. Swets JA (1986) Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance. *Psychol Bull* 99: 181–198.

26. Yaniv I, Yates JF, Smith JEK (1991) Measures of discrimination skill in probabilistic judgment. *Psychol Bull* 110: 611-617.

27. Yates JF (1990) *Judgment and Decision Making* (Prentice Hall, Englewood Cliffs, NJ), 430 pp.

28. Cox DR (1958) Two further applications of a model for binary regression. *Biometrika* 45: 562-565.

29. Budescu DV, Johnson TR (2011) A model-based approach for the analysis of the calibration of probability judgments. *Judgment and Decision Making* 6: 857-869.

30. Lichtenstein S, Fischhoff B (1977) Do those who know more also know more about how much they know? *Organ Behav Hum Perform* 20: 159-183.

31. Karvetski CW, Olson KC, Mandel DR, Twardy CR (2013). Probabilistic coherence weighting for optimizing expert forecasts. *Decis Anal* 10: 305-326.

32. Turner BM, Steyvers M, Merkle EC, Budescu DV, Wallsten TS (in press) Forecast aggregation via recalibration. *Mach Learn*. doi: 10.1007/s10994-013-5401-4

33. Shlomi Y, Wallsten TS (2010) Subjective recalibration of advisors' probability estimates. *Psychon Bull Rev* 17: 492-498.

34. Pearce J, Ferrier S (2000). Evaluating the predictive performance of habitat models developed using logistic regression. *Ecol Modell* 133: 225-245.

35. Akaike H (1974) A new look at the statistical model identification. *IEEE Trans Automat Contr* 19: 716–723.

36. Koehler DJ, Brenner L, Griffin D (2002) in Heuristics and Biases: The Psychology of Intuitive Judgment, eds Gilovich T, Griffin D, Kahneman D (Cambridge University Press, Cambridge, U.K.) pp. 686-715.

37. Fischhoff B, Bruine de Bruin W (1999) Fifty-fifty=50%? *J Behav Decis Mak* 12: 149-163.

38. Karmarkar US (1978) Subjectively weighted utility: A descriptive extension of the expected utility model. *Organ Behav Hum Perform* 21: 61-72.

39. Tetlock PE, Kim JI (1987) Accountability and judgment processes in a personality

prediction task. *J Pers Soc Psychol* 52: 700-709.

40. Tetlock PE (1985) Accountability: A social check on the fundamental attribution error. *Soc Psychol Q* 48: 227-236.

41. Chaiken S (1980) Heuristic versus systematic information processing and the use of source versus message and cues in persuasion. *J Pers Soc Psychol* 39: 752-766.

42. Hagafors R, Brehmer B (1983) Does having to justify one's decisions change the nature of the judgment process? *Organ Behav Hum Perform* 31: 223-232.

43. Arkes HR, Kajdasz J (2011) in *Intelligence Analysis: Behavioral and Social Scientific Foundations,* eds Fischhoff B, Chauvin C (National Academies Press, Washington), pp 143-168.

44. Kadane J, Fischhoff B (2013) A cautionary note on global recalibration. *Judgment and Decision Making* 8: 25-28.

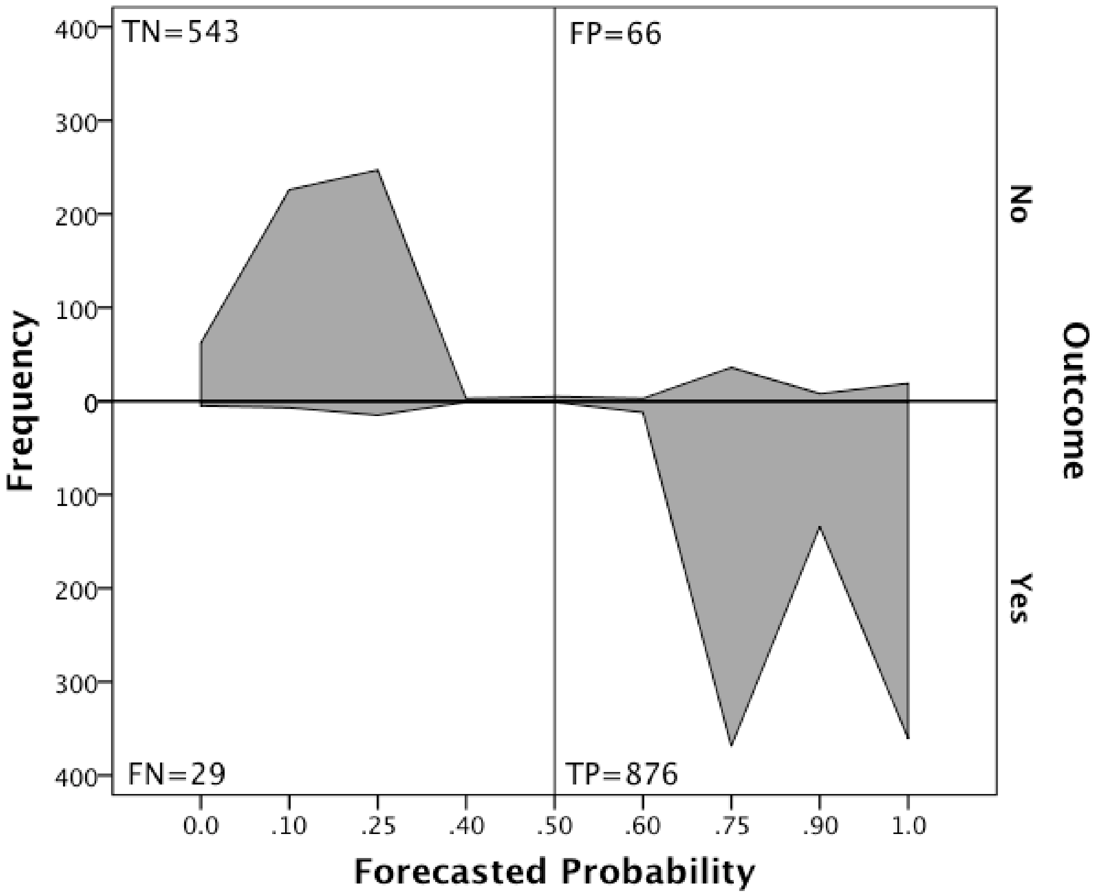
Figure Legends

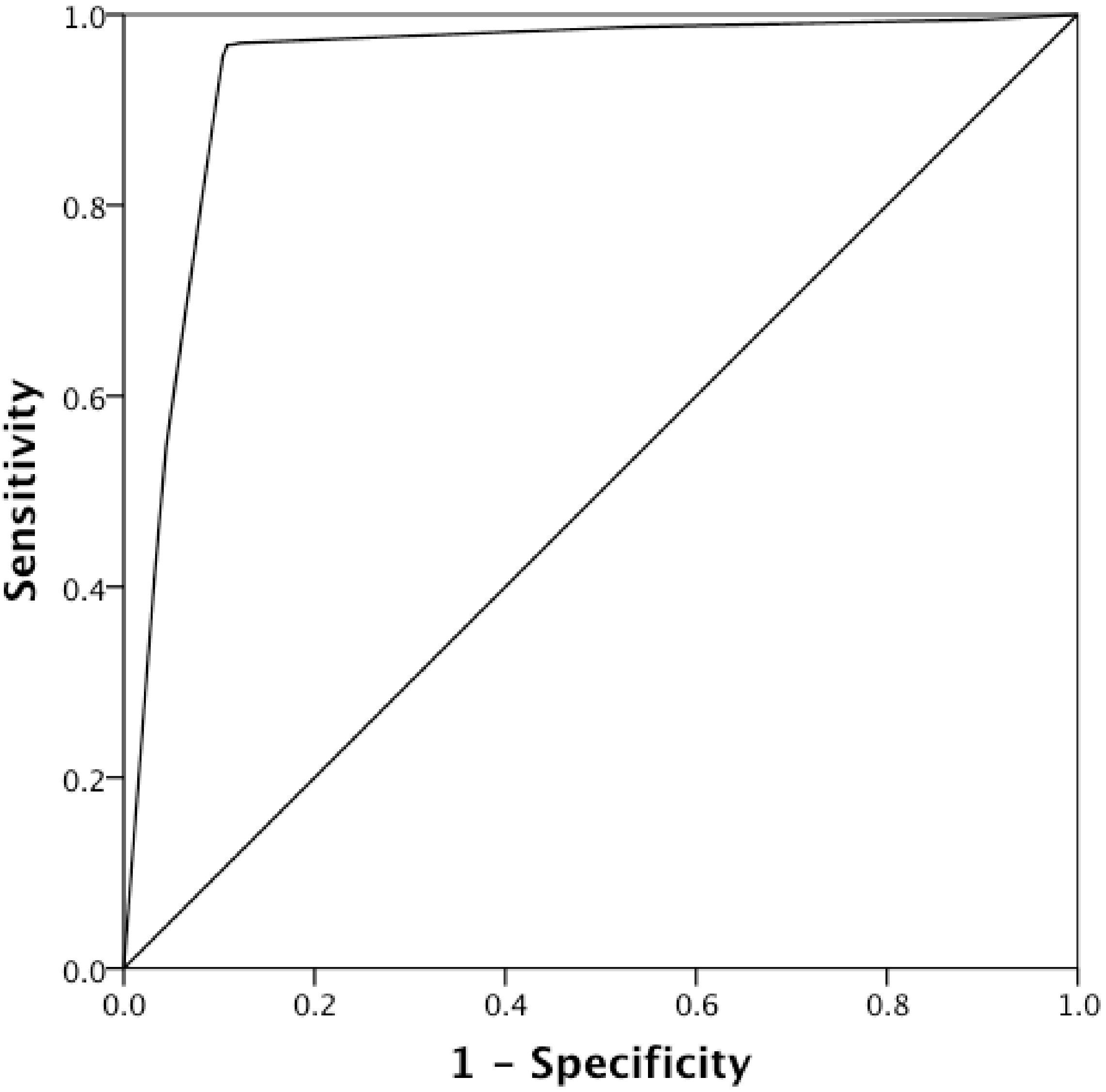
Fig. 1. Discrimination diagram. TN = true negatives; FP = false positives; FN = false negatives; TP = true positives.

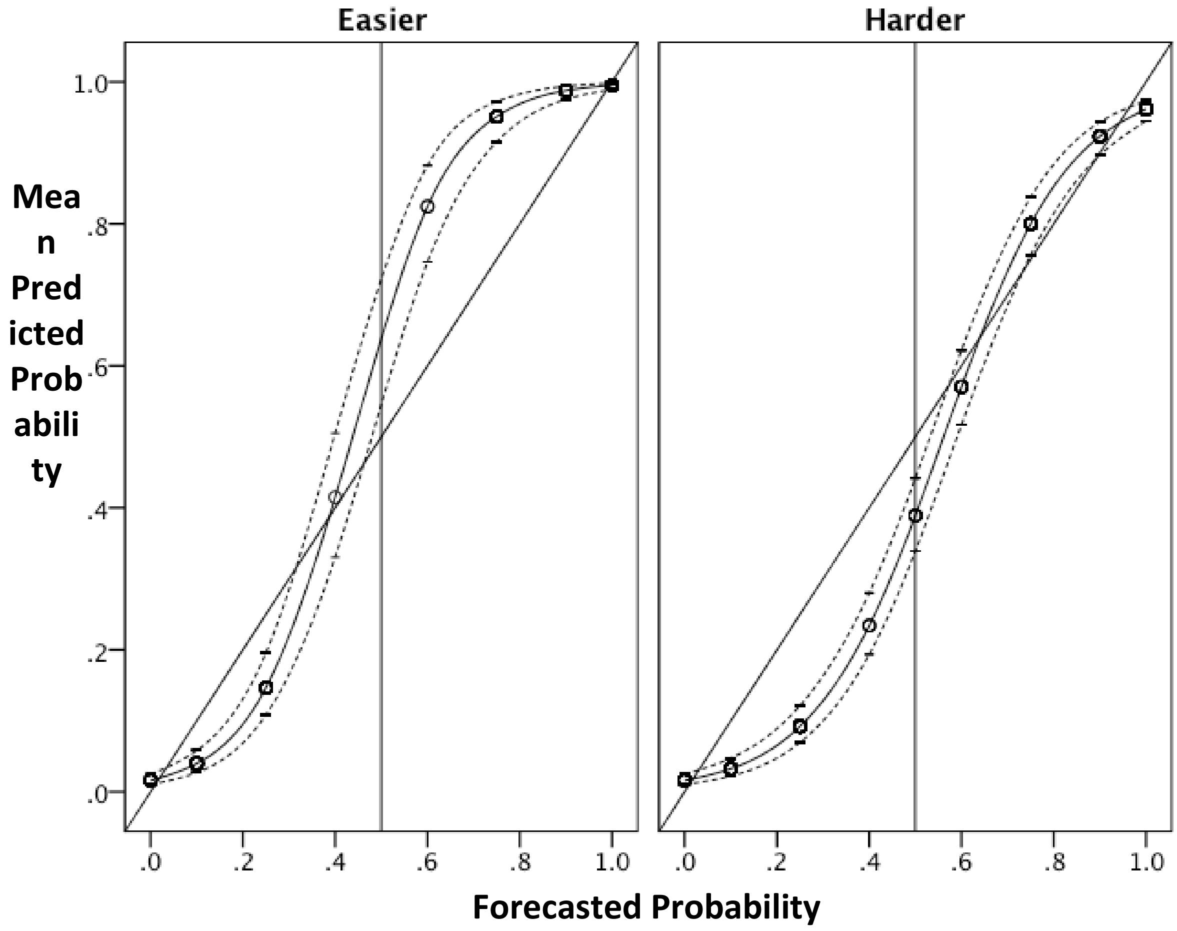
Fig. 2. ROC curve.

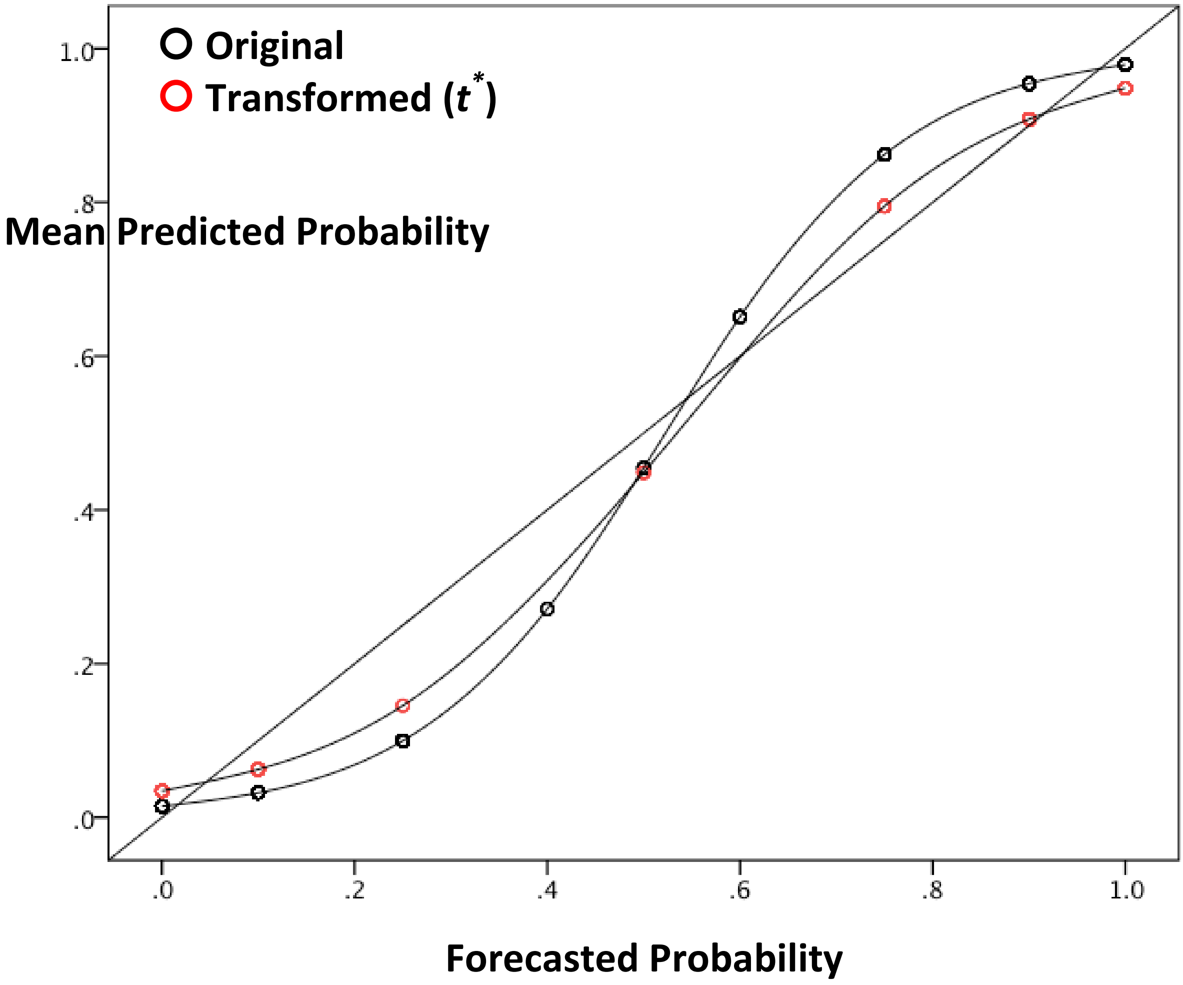
Fig. 3. Model-based calibration curves. Dotted lines show 95% confidence intervals.

Fig. 4. Calibration curves before and after recalibration to t^*.


TN=543
FP=66
FN=29
TP=876
Frequency
400
300
200
100
0
100
200
300
400
0.0
.10
.25
.40
.50
.60
.75
.90
1.0
Forecasted Probability
Outcome
No
Yes


1.0
0.8
0.6
0.4
0.2
0.0
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1 - Specificity


Easier
Harder
Mean Predicted Probability
1.0
.8
.6
.4
.2
.0
.0
.2
.4
.6
.8
1.0
Forecasted Probability


Original
Transformed (t^*)
Mean Predicted Probability
1.0
.8
.6
.4
.2
.0
.0
.2
.4
.6
.8
1.0
Forecasted Probability

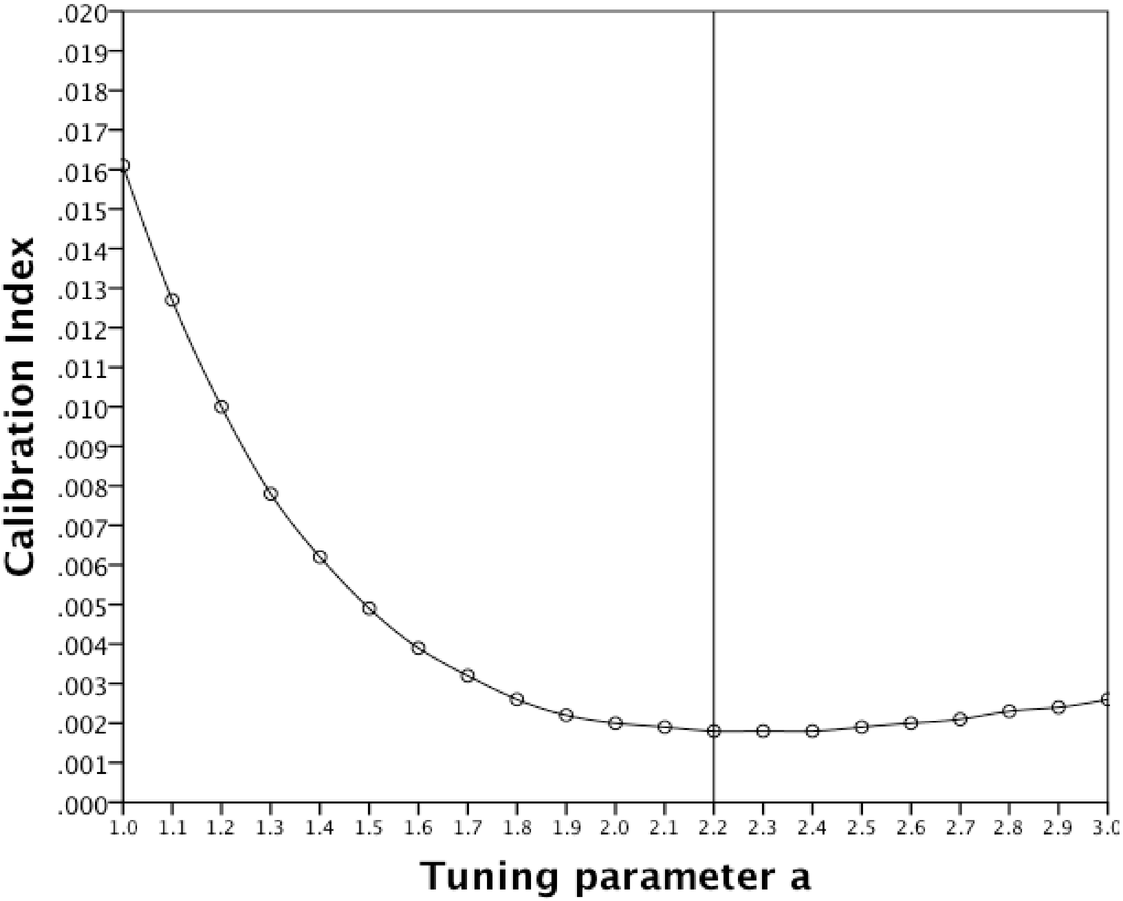


Fig. S1. Calibration index value as a function of tuning parameter, *a*, in Karmarkar transformation, *t*.

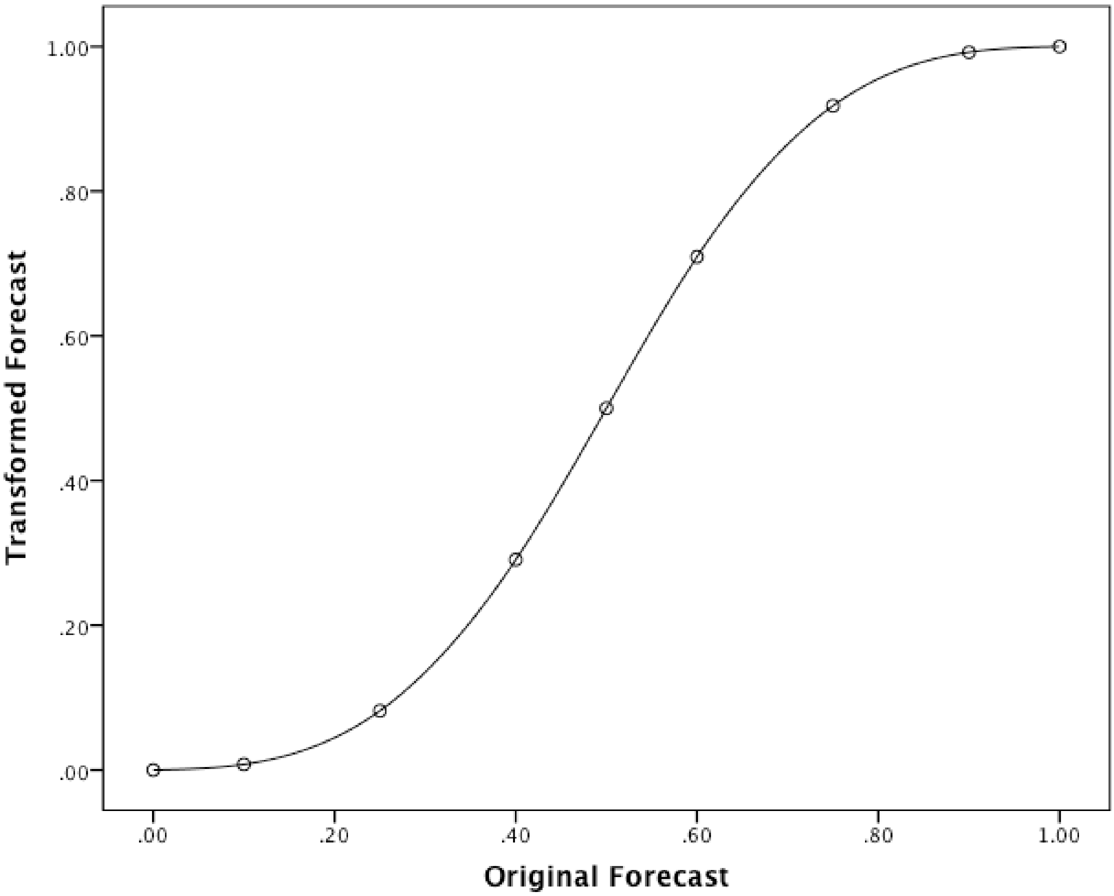


Fig. S2. Transformed forecast, *t*, as a function of original forecast, f_k.

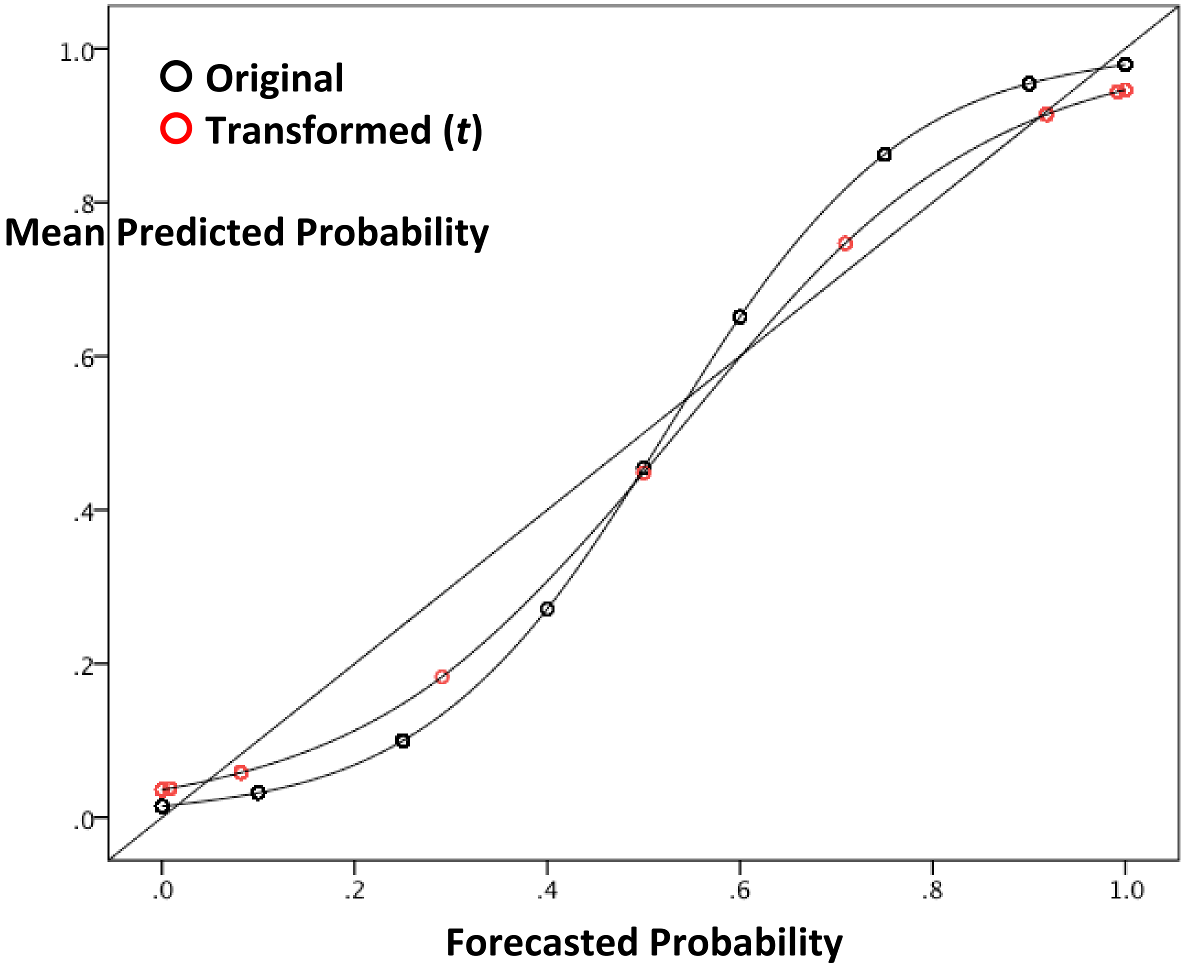


Fig. S3. Calibration curves before and after recalibration to *t*.