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

DEFENCE R&D DÉFENSE

A Quantitative Assessment of the Quality of Strategic Intelligence Forecasts

David R. Mandel
Alan Barnes
Karen Richards

Defence R&D Canada
Technical Report
DRDC Toronto TR 2013-036
March 2014

Canada

A Quantitative Assessment of the Quality of Strategic Intelligence Forecasts

David R. Mandel
Alan Barnes
Karen Richards

Defence R&D Canada, Toronto

Technical Report
DRDC Toronto TR 2013-036
March 2014

Principal Author

Original signed by David R. Mandel

Dr. David R. Mandel

Defence Scientist, Socio-Cognitive Systems Section

Approved by

Original signed by Keith Stewart

Keith Stewart

Section Head, Socio-Cognitive Systems Section

Approved for release by

Original signed by Joseph V. Baranski

Joseph V. Baranski

Chair, Knowledge and Information Management Committee; Chief Scientist

In conducting the research described in this report, the investigators adhered to the policies and procedures set out in the *Tri-Council Policy Statement: Ethical conduct for research involving humans* (2010) as issued jointly by the Canadian Institutes of Health Research, the Natural Sciences and Engineering Research Council of Canada and the Social Sciences and Humanities Research Council of Canada.

Abstract

This report describes a field study of the quality of probabilistic forecasts made in Canadian strategic intelligence reports. The researchers isolated a set of 1,422 probabilistic forecasts from intelligence memoranda and interdepartmental committee reports for which outcome information about the forecasted events was available. These data were used to study forecast quality measures, including calibration and discrimination indices, commonly employed in other areas of expert judgment monitoring research (e.g., meteorology or medical diagnosis). Predictions were further categorized in terms of other variables, such as the organizational source, forecast difficulty, and forecast importance. Overall, the findings reveal a high degree of forecasting quality. This was evident in terms of calibration, which measures the concordance between probability levels assigned to forecasted outcomes and the relative frequency of observed outcomes within that assigned category. It was also evident in terms of adjusted normalized discrimination, which measures the proportion of outcome variance explained by analysts' forecasts. The main source of bias detected in analytic forecasts was underconfidence: Analysts often rendered forecasts with greater degrees of uncertainty than were warranted. Implications for developing outcome-oriented accountability systems, adaptive learning systems, and forecast optimization procedures to support effective decision-making are discussed.

Résumé

Le rapport traite d'une étude pratique sur la qualité des prévisions probabilistes dans les rapports de renseignements stratégiques canadiens. Les chercheurs ont retenu une série de 1 422 prévisions probabilistes effectuées dans des notes de renseignements et des rapports de comités interministériels pour lesquelles de l'information sur les situations réelles était disponible. Ils ont utilisé les données pour examiner les mesures de la qualité des prévisions, dont les indices d'étalonnage et de discrimination, couramment employés dans d'autres domaines de recherche et de surveillance où des experts portent des jugements (p. ex. météorologie, diagnostic médical). Ils ont en outre classé les prévisions en fonction d'autres variables, comme leur provenance (organisation), leur niveau de difficulté et leur importance. Dans l'ensemble, les conclusions révèlent que la qualité des prévisions est élevée. L'étalonnage, qui mesure la concordance entre le niveau de probabilité attribué aux résultats prévus et la fréquence relative des résultats observés dans la catégorie visée, l'a confirmé de manière évidente. Il en va de même de la discrimination normalisée et ajustée, qui sert à mesurer la proportion de la variance des résultats expliquée par les prévisions des analystes. La principale source de biais est la confiance insuffisante des analystes à l'égard de leurs prévisions probabilistes : elles comportaient souvent des degrés d'incertitude plus élevés que nécessaire. Le rapport traite également de ce qu'il faudra pour mettre au point des systèmes et procédures qui contribueront à l'efficacité des prises de décisions – systèmes de reddition de comptes fondés sur les résultats, systèmes d'apprentissage adaptatif et procédures d'optimisation des prévisions.

This page intentionally left blank.

Executive summary

A Quantitative Assessment of the Quality of Strategic Intelligence Forecasts

David R. Mandel, DRDC Toronto; Alan Barnes, Intelligence Assessment Secretariat of the Privy Council Office; Karen Richards, DRDC Toronto
TR 2013-036; Defence R&D Canada, Toronto; March 2014.

Introduction: This report describes a field study of the quality of probabilistic forecasts made in Canadian strategic intelligence reports. The researchers examined a set of 1,422 forecasts from intelligence memoranda and interdepartmental committee reports for which outcome information about the forecasted events was available. These data were used to study forecast quality measures, including calibration and discrimination indices, commonly employed in other areas of expert judgment monitoring research (e.g., meteorology or medical diagnosis). Forecasts were further categorized in terms of other variables, such as the use of numeric or verbal probabilities, the organizational source of the forecast, forecast difficulty, and forecast importance.

Results: Overall, the findings reveal a high degree of forecast quality. This was evident in terms of calibration, which measures the concordance between probability levels assigned to forecasted outcomes and the relative frequency of observed outcomes within that assigned category. It was also evident in terms of adjusted normalized discrimination, which measures the proportion of outcome variance explained by analysts' forecasts. The main source of bias detected in this study was underconfidence in probabilistic forecasting: Analysts often rendered forecasts with greater degrees of uncertainty than were warranted. Evidently, analysts do not fully appreciate the extent of their success in correctly classifying world events that will occur from those that will not.

Significance: The methods described in this report provide a clear example of how intelligence organizations could objectively and systematically track various facets of their forecasting capability. As such, they provide an example of how outcome-based accountability measures could be implemented to gauge analytic accuracy. The results themselves are significant given that they reveal for the first time the level of forecast accuracy of a large corpus of real strategic intelligence judgments. The findings indicate a high level of forecasting accuracy. The finding that inaccuracy was largely attributable to underconfidence is important for at least two reasons. First, it suggests that the primary source of error in analytic forecasts is correctable by "recalibrating" the judgments to make them less conservative. Second, it raises questions about the effectiveness of current analytic training practices that warn analysts of the pitfalls of overconfidence. At a minimum, the findings suggest that analysts in training should be taught to consider the costs of various types of errors or biases, including opposing ones like over- and under-confidence.

A Quantitative Assessment of the Quality of Strategic Intelligence Forecasts:

David R. Mandel, DRDC Toronto; Alan Barnes, Intelligence Assessment Secretariat of the Privy Council Office; Karen Richards, DRDC Toronto
TR 2013-036; Defence R&D Canada, Toronto; March 2014.

Introduction ou contexte : Le rapport traite d'une étude pratique sur la qualité des prévisions probabilistes dans les rapports de renseignements stratégiques canadiens. Les chercheurs ont examiné une série de 1 422 prévisions effectuées dans des notes de renseignements et des rapports de comités interministériels pour lesquelles de l'information sur les situations réelles était disponible. Ils ont utilisé les données pour examiner les mesures de la qualité des prévisions, dont les indices d'étalonnage et de discrimination, couramment employés dans d'autres domaines de recherche et de surveillance où des experts portent des jugements (p. ex. météorologie, diagnostic médical). Ils ont en outre classé les prévisions en fonction d'autres variables – si elles étaient des probabilités chiffrées ou non chiffrées par exemple –, comme leur provenance (organisation), leur niveau de difficulté et leur importance.

Résultats : Dans l'ensemble, les conclusions révèlent que la qualité des prévisions est élevée. L'étalonnage, qui mesure la concordance entre le niveau de probabilité attribué aux résultats prévus et la fréquence relative des résultats observés dans la catégorie visée, l'a confirmé de manière évidente. Il en va de même de la discrimination normalisée et ajustée, qui sert à mesurer la proportion de la variance des résultats expliquée par les prévisions des analystes. Selon l'étude, la principale source de biais est la confiance insuffisante des analystes à l'égard de leurs prévisions probabilistes : elles comportaient souvent des degrés d'incertitude plus élevés que nécessaire. De toute évidence, les analystes mésestiment grandement leur capacité à déterminer les évènements mondiaux qui se produiront et ceux qui n'auront pas lieu.

Importance: Les méthodes présentées dans le rapport constituent de très bons exemples de la façon dont les services de renseignements pourraient suivre objectivement et systématiquement diverses facettes de leurs capacités de prévisions. Ainsi, elles montrent comment il serait possible d'adopter des mesures de reddition de comptes fondées sur les résultats pour évaluer l'exactitude des analyses. Les résultats mêmes du rapport sont importants du fait qu'ils révèlent pour la première fois le degré d'exactitude d'un grand corpus de conclusions réelles d'analystes du renseignement stratégique. Selon ces résultats, l'exactitude des prévisions est très grande. La conclusion selon laquelle l'inexactitude était surtout attribuable à la confiance insuffisante à l'égard des prévisions est importante pour au moins deux raisons. Tout d'abord, elle semble indiquer qu'il est possible de corriger la source principale d'erreur dans les prévisions des analystes en « rajustant » les jugements de ces derniers pour les rendre moins prudents. Ensuite, elle soulève des questions quant aux effets de la pratique actuelle selon laquelle les apprentis analystes, au cours de leur formation, sont mis en garde contre les pièges posés par une trop grande confiance. À tout le moins, les conclusions semblent indiquer que les analystes en formation devraient apprendre à tenir compte des coûts de divers types d'erreur ou de biais, dont certains qui s'opposent, comme la trop grande confiance et la confiance insuffisante.

Table of contents

List of figures

List of tables

Acknowledgements

This research was supported in part by funding from Applied Research Program Project 15dm entitled "Understanding and Augmenting Human Analytic Capabilities," under the direction of David Mandel (first author) and under the sponsorship of the Director General of Intelligence Capability, Chief of Defence Intelligence. We thank Rex Brynen and John Hannigan for their diligence and expertise in serving as coders in the study described in this report, and we thank James Moore for constructive feedback on an earlier draft of this report.

This page intentionally left blank.

1 Introduction

An important function of strategic intelligence is to help decision makers better understand the factors shaping the world around them and to improve their ability to anticipate future events. Towards that end, analysis that offers accurate and timely forecasts about consequential events can be of significant value to decision makers. Not all intelligence is predictive, even in a probabilistic sense, but forecasting and prediction are an important part of intelligence analysis, especially strategic intelligence, which must help decision makers anticipate surprises and significant developments over a longer view.

The importance of forecasting in intelligence assessment naturally ought to raise the question of how good analysts' forecasts actually are. Knowing how good analysts are at forecasting events would seem to be a critical piece of information for various stakeholder constituencies. First of all, this would seem to be critical for consumers to gauge, given that they rely on intelligence to inform and, ultimately, improve their planning and decision-making processes. How are they to know how much trust in the analytic enterprise is warranted unless there is some objective measure of the analytic track record to consult?

Likewise, managers of intelligence analysts, who presumably want to be able to objectively account for the quality of assessments under their control, should want to know how well their analysts are at forecasting events of real or potential consequence. An objective record may be even more important for managers than consumers since they are formally responsible for quality control within their remit. In some sense, if they were systematically monitoring performance and providing clear feedback and corrective advice to analysts, consumers might not have to wonder how good the intelligence they are receiving actually is.

At the highest levels of intelligence directorship, managers could benefit from an objective score card as a buffer against the sort of "accountability ping-pong" that Tetlock and Mellers (2011a) describe in which the intelligence community (IC) shifts its tolerance for false-positive and false-negative errors depending on the nature of the most recent failure to "get it right." After a false negative, such as failing to detect 9/11, accountability pressures push the IC towards a less tolerant (beta) criterion for false positives, which increases the risk of future false positives, such as the US IC's erroneous conclusions regarding Saddam Hussein's WMD program. While objective measures of the IC's performance will not eliminate accountability ping-pong, it could attenuate or help rationalize the shifts when they occur. In particular, such metrics should help the IC implement Betts's (2007) recommendation that criticism of the IC and subsequent reform attempts shift away from an unrealistic "zero defects" standard of performance and towards an "improved batting average" goal. It is hard to improve one's batting average without a scorecard or even scoring rules. As Betts notes, "there are no clear indicators of the ratio of failure to success in intelligence" (p. 21).

Of course, analysts themselves have a vested interest in knowing how good their forecasts really are. Part of that interest ought to stem from natural curiosity, the same sort of curiosity one has about their level of performance after taking a challenging test—one wants to know at least how well one *scored*, if not how well one *ranked* among one's cohort. Having objective measures of performance can also serve as a valuable learning aid, revealing to analysts just how well they actually perform, how and where they went wrong, and perhaps how they might do better in the future (e.g., see Rieber, 2004, on the benefits of calibration training). Without such feedback, analysts simply cannot know how good their track record of forecasts actually is—at least, not in an objective and systematic sense.

Ultimately, all citizens have a vested interest in the quality of intelligence assessments, predictive or otherwise. Maintaining a state's intelligence apparatus is a costly enterprise. The US National Intelligence Program was US$53.1 billion in fiscal 2010 and US$1.5 billion more in fiscal 2011. Military intelligence added US$27 billion to the 2010 budget, for a total of US$80.1 billion (see "Fiscal 2011 U.S. intelligence budget," 2011). While Canada's intelligence budget is a fraction of that cost, the cost nevertheless should be warranted. Intelligence organizations are also tasked with vital national security duties, which are presumably correlated to some extent with citizens' personal security. The public has the right to know how well the IC is carrying out those duties, provided that such information is conveyed in a manner that itself does not threaten national security.

In short, then, it appears there should be multiple pressures for objective verification of intelligence product quality in general and forecast quality in particular. Remarkably, however, intelligence forecasts have not been subjected to careful, long-term evaluation using objective methods that provide meaningfully quantifiable measures of forecast quality. The work described later in this report is, in fact, not only rare but unique in this regard. It represents, to the best of our knowledge, the first systematic, long-term evaluation of the quality of analytic forecasts extracted from real intelligence reports, using the same objective measures of forecasting quality—Brier scores and their component indices of calibration and discrimination (Brier, 1950; Murphy, 1973; Yaniv, Yates, & Smith, 1991)—that have been used to evaluate human performance in other expert forecasting domains (e.g., Åstebro & Koehler, 2007; Goodman-Delahunty, Granhag, Hartwig, & Loftus, 2010; Lin & Bier, 2008; Murphy & Winkler, 1984) and among non-experts in basic research on human judgment (e.g., Erev, Wallsten, & Budescu, 1994; Lichtenstein, Fischhoff, & Phillips, 1982; Moore & Healy, 2008).

There are, of course, plausible reasons why intelligence organizations have not engaged in this sort of performance evaluation activity. For one thing, such exercises require certain types of data that are usually unavailable. In order to keep score of analytic forecasting performance, the outcomes of predicted events need to be tracked and recorded in a systematic manner. If such tracking were to be carried out by intelligence analysts or even their managers, then they may have to justify the allocation of their resources to this sort of activity. Why not simply put the time required into tracking past outcomes of predicted events into more analysis that might help decision makers with pressing policy or command decisions requiring immediate or future action? That is, they might have to account for their accountability efforts, and that might serve as a deterrent.

The measures of calibration and discrimination used in this report also require numeric probabilities to be associated with forecasted events. If such probabilities are not actually provided by the analysts making the predictions, then they would need to be inferred from their use of verbal expressions of uncertainty—terms such as "likely," "probably not," and so forth. That is a subjective exercise even under the most rigorous conditions because it requires the forecaster or a third-party assessor to judge what a probabilistic judgment rendered in verbal terms would have meant had it been rendered in numeric terms instead. Most intelligence analysts do not issue forecasts using numeric probabilities, and few intelligence organizations have standards established that would provide a reliable basis for inferring such numbers from verbal statements. Thus, a reliance on verbal expressions of uncertainty in intelligence forecasts poses a barrier to objective, outcome-based performance measurement. That barrier, on its own, is not insurmountable since there are statistical methods for examining the quality of forecasts made on an ordinal probability scale, such as would be the case if a set of verbal probability terms such as "very unlikely," "unlikely," "even chance," "likely," and "very likely" were used instead of numerical forecasts. Somers *d* statistic provides one example and has been discussed in this context by Liberman and Tversky (1993).

The fact that forecasts are not offered in numeric terms perhaps reveals a more deep-seated set of reasons why objective, quantifiable measures of forecasting ability have not been implemented. There is a suspicion of quantification in the intelligence community, at least when it comes to forecasting or prediction. Many analysts, managers, and even educators believe that predictions should not be given in numeric terms because that would convey a false sense of precision or falsely communicate that the estimate was scientific. Those who claim the latter as an argument against quantification typically also like to point out that (in their view) intelligence analysis is an art, not a science. The intelligence community as a whole is also not well versed in probability theory, and there is little understanding of the different senses of probability, especially how subjective probabilities can be used to quantify degrees of belief in the veracity of a hypothesis or prediction.

Indeed, in 2011, the first author attended an intelligence workshop in which a senior manager of an intelligence organization serving as a workshop panellist claimed that intelligence is not in the business of making predictions. When challenged on that point by audience members, the manager explained that intelligence could not offer predictions because predictions required having access to all the pertinent information so that a deterministic "this will happen with absolutely certainty" assessment could be made. Evidently, the manager's understanding of the term "prediction" did not include a probabilistic concept. Prediction was more or less akin to clairvoyance: saying with 100% certainty what will happen. While predictions or forecasts can be issued with complete certainty, they surely do not have to be. Most intelligence judgments, as Sherman Kent (1964) noted long ago, are qualified by words that vaguely convey the level of probability or certainty associated with the assessment.

Putting aside such gross misconceptions (e.g., prediction = clairvoyance), the arguments against quantifying uncertainty are flimsy and should be easy to defeat. If numeric forecasts offered as point estimates are overly precise, they could be offered along with confidence intervals. For instance, an analyst could say, "we estimate that there is a 75% chance of event *x* occurring in time frame *y*, and we are 90% confident that the true probability lies between 60% and 90%." Instead, using verbal terms, the analyst might issue a vague and/or ambiguous prediction, such as "we believe that it is (highly) likely that event *x* will happen in time frame *y*." Consumers of such an estimate would not only have to deal with the uncertainty of the estimate itself, as in the numeric case, they would also have to deal with the uncertainty of what terms like "likely" mean—what such terms mean to them and what they might have meant to the analyst issuing the vague prediction. That hardly seems like a good solution to fair but addressable concerns about over-precision.

Resistance to using numeric probabilities in forecasts also severely diminishes the potential for aggregating judgments. Imagine that a decision maker would like to know how likely it is that *x* will happen, where *x* may represent a given threat scenario. For *x* to happen, let us further imagine that three preconditions (*a*, *b*, and *c*) need to be met. To simplify the example, let us assume that the three preconditions are jointly sufficient and individually necessary for *x*. Thus, the probability of *x* is the product of the probabilities of the preconditions:

$$P(x) = P(a) \times P(b) \times P(c)$$

If the probabilities of the preconditions were expressed numerically, the estimation of $P(x)$ would be a simple matter of arithmetic. In contrast, the use of verbal expressions of uncertainty precludes such aggregation or turns it into a virtually meaningless exercise. Imagine, for instance, that the three preconditions are forecasted as having a "remote chance," an "even chance," and

being "almost certain," respectively.[1] How would one be able to combine those terms to get the best estimate of the probability that *x* will occur? One could not, and one would probably be better off not trying because the resulting estimate would, in all likelihood, be highly misleading. Rather than multiplying through the terms, an intuitive estimation would more likely approximate an averaging of those terms. This is because people are well adapted to compute an average representation of a set of related items, but they find it quite difficult to intuitively sum or multiply the same set of items. This generally leads to an overestimation of the probability of conjunctions with at least one very improbable element (Mandel, 2008b; see also Kahneman, 2011).

To illustrate this, imagine that the numeric estimates for the preconditions in our earlier example had been estimated at a 1%, 50%, and 99% chance of occurring, respectively. If so, then the best estimate of the probability of *x* is .00495—roughly 5/1000 or half of a 1% chance. But, if one erroneously averaged these values instead, one would be prone to mistakenly conclude that the average was closer to 50%, a difference of two orders of magnitude.

Moreover, with confidence intervals on the precondition probabilities, more sophisticated aggregative estimates could be easily generated, including high and low probability estimates for the threat scenario *x*. Quantification of the estimates would also facilitate the use of if-then analyses, which could show how a compound probability would change depending on changes to one or more precondition probabilities. For instance, using numbers, it becomes clear that small probabilities are particularly influential in the estimation of compound probabilities. For instance, if the lowest estimate were raised by five percentage points, the revised estimate of $P(x)$ would be approximately .03. In contrast, had the highest estimate been lowered by five percentage points, the revised estimate of $P(x)$ would have been .0047. Thus, whereas the former adjustment led to a six-fold increase in estimated probability, the latter adjustment by the same amount (five percentage points) led to no change when rounded to the nearest tenth of a percent. In short, there are many distinct analytic advantages to quantifying uncertainty, and that applies to the quantification of imprecise measures of uncertainty.

Briefly addressing the quip that quantification ought to be rejected because analysis is more like an art than a science, we wonder precisely what objectors to quantification might mean by this if, in fact, they were serious about defending the claim. If analysis is artful, in the sense of being a skilled and imaginative enterprise, then its distinction with science is, at once, false and irrelevant. Good scientists must, of course, also be skilful and imaginative; but, in any case, these qualities neither confirm nor deny a requirement for quantification. It is beside the point.

Alternatively, perhaps art is meant by objectors of quantification to refer to the fact that artists embrace subjectivity. If two people experience a painting differently, that is fine. Science, in contrast, seeks objectivity and verification against ground truth. One hopes that proponents and opponents of a given scientific theory can nonetheless agree whether the findings of a well-conceived empirical test confirm or disconfirm one of the theory's hypotheses. However, if the objection to quantification rests on the claim that intelligence analysis should be more like making art than conducting science, then we would strongly disagree with the proposition. Unlike art, which has no ground truth, analytic assessments can eventually be evaluated in light of ground truth. Forecasts can be verified and, if they were clear enough, different parties should be able to agree on which forecasts were accurate and which were not. Indeed, intelligence analysis also shares important process elements with the scientific method, the most obvious of which is the cornerstone of providing fair, unbiased tests of hypotheses with the aim of establishing the

[1]Some readers may recognize that these terms are, in fact, ones that have been advocated for use in past National Intelligence Estimates of the (U.S.) National Intelligence Council (for a discussion of the use of these terms, see Kesselman, 2008).

truth. This is why analysts learn structured analytic techniques like Heuer's (1999) Analysis of Competing Hypotheses (ACH) and others meant to "debias" hypothesis testing.[2]

In any case, perhaps a better metaphor is intelligence as a sport. Like sports, intelligence takes place within a competitive environment. Intelligence organizations play for a team—their state. They contribute to the team's goals—national security and furthering the state's prosperity. They have opponents—other states or organizational entities that are adversaries or enemies. The various teams employ strategies and tactics to win. And so on. When it comes to quantification, however, the analogy currently breaks down. Most sports have clear ways of tallying the scorecard. One might disagree with a referee or umpire, but the rules for scorekeeping—that is, for quantifying performance—are clear. Scorekeeping for intelligence is, of course, a much more daunting task. Some important functions might not be amenable to quantifiable and objective scorekeeping. In those areas, subjective evaluations by consumers or managers or even self-evaluations might be the best one can realistically achieve. In other areas, however, progress towards objective scorekeeping could be made but is not (for a related discussion, see McClelland, 2011). The lack of progress in those areas reflects other barriers—a lack of will or knowledge of how to implement objective assessment processes in intelligence organizations.

This report presents the findings of a long-term effort to implement a more objective, proactive, and readily quantifiable method of scorekeeping for intelligence forecasts. The primary aim of this report is to present a detailed exposition of the findings, one that is much more comprehensive than what would be permitted within the confines of a normal-length journal article. Another aim of the report is to provide a detailed written record of how this work came about and how it was executed—a task that similarly requires more space than a typical journal article would afford. We also aim to provide a level of detail regarding the performance metrics used here that should allow an educated non-expert in these methods to find the subsequent analyses understandable.

Although the findings presented here could be juxtaposed with other studies of expert prediction, used to test theoretical propositions, or discussed in terms of their implications for implementing outcome-based accountability processes within intelligence organizations (e.g., see Tetlock & Mellers, 2011b), that is not our aim in this report. While striving to be comprehensive in the exposition of the methods and results, this report is not intended to replace the dissemination of key findings and their basic and applied significance in peer-reviewed journals. We anticipate the subsequent release of such publications and expect that they may rely on the present report for background information. Ultimately, our aim here is to disseminate a full descriptive record of the findings early on so that they may be shared with key stakeholders in the Canadian and allied defence and security community.

[2]Ironically, as a recent report on field research in the context of intelligence and counter-intelligence (National Research Council, 2010) makes clear, whether such techniques actually debias analysts or analytic products in any significant manner is a topic that has received little scientific attention in applied research. More recently, Greenwald (2012) offers a pessimistic assessment of the success with which theoretical debates in science are actually resolved through a "competing hypotheses" approach.

2 Evolution of the study

In 2007, a fortuitous meeting was arranged, which subsequently led to the partnership that enabled the present study. DRDC Toronto had recently undergone a science and technology (S&T) capability and program review and subsequent centre re-organization. One of its new sections (the now defunct Adversarial Intent section) was exploring opportunities to conduct research and development (R&D) in support of clients in the Canadian intelligence community.

In 2007, several defence scientists from the section, along with the section head, met with potential intelligence stakeholders in Ottawa to explore partnership opportunities. A meeting was arranged with some of the directors and analysts from the International Assessment Staff (IAS) of the Privy Council Office (PCO).[3] The meeting was introductory. However, it emerged at that meeting that one of the scientists (David Mandel) had expertise in the area of human judgment and decision making and that one of the directors (Alan Barnes), who headed the Middle East and Africa (MEA) Division within the IAS, was engaged in an effort to monitor the quality of his division's predictive judgments.

Barnes arranged a second meeting in Ottawa soon after, where he presented the method and preliminary results of his judgment-quality monitoring activity. The use of numerical probabilities in draft reports by the MEA Division was initially intended as a means of improving the transparency of judgments during internal discussions. Once the Division had amassed a number of assessments with numerical probabilities, it became clear that this could also be used to assess the quality of the judgments that had been made. Mandel realized that Barnes had effectively established the conditions for a quantitative study of analytical forecasting success, but that he had not yet applied well-established measures of forecast quality, such as calibration and discrimination indices, to the data. From that point on, the two agreed to work together, with Mandel taking charge of the quantitative analysis. The partnership was later formalized through a February 12, 2008, Memorandum of Understanding (MOU) between the Department of National Defence as represented by DRDC Toronto and PCO as represented by IAS. A renewal of that MOU was signed in April 2012.

On April 1, 2008, Mandel became the principal investigator of a DRDC Applied Research Program project entitled "Understanding and Augmenting Human Analytic Capabilities," sponsored by the Director of Intelligence Capability (now the Director General of Intelligence Capability) within Chief of Defence Intelligence (CDI). The work reported here was formally taken on as an element of that project, and the initial effort undertaken by Barnes prior to forging a partnership with Mandel was extended, in part, through funding from that project.

This report summarizes the results of this study up to the end of its second phase. At this point, there is a significant need to capture the findings from Phases 1 and 2 in a written report. That need is mainly driven by three factors.

First, the sheer quantity of analyses produced up to this point must be organized and put into an easily accessible form with a table of contents. Until now, results have been sent between the research team at DRDC Toronto and Barnes at IAS (as the primary stakeholder) as attachments via electronic mail, along with explanatory notes. While this approach has allowed new results to be rapidly shared, it is insufficient as a cumulative repository of the findings.

[3] The International Assessment Staff has since reverted to its former name of Intelligence Assessment Secretariat and retains the same abbreviation.

Second, the methodological description of the study must be properly captured. With the passage of time, there is an incremental risk of critical information loss as people who have worked on the study move on to other posts, retire, or simply forget. The process of drafting this report thus serves as a check on ensuring that all the methodological questions that we pose to ourselves are fully answered.

Finally, there is a need to disseminate the findings in a written form because of an increasing number of requests from others to reference our work. The first phase of the study has been reported at several professional meetings (Mandel, 2008a, 2009a, 2009b, 2009c), and the first and second phases have now been reported as well (Mandel, 2011a, 2011b). These presentations have generated several requests for a citable, written report of the findings, which, until now, has been unavailable. As a result, extant citations of the present study (e.g., Arkes & Kajdasz, 2011; National Research Council, 2011, chapters 2, 3, and 7) are to conference presentations, which, even if accompanied by written notes, are much harder to obtain than a published report. In the case of one recent book (Gardner, 2010), there is a substantial discussion of the present study based on an interview conducted by the book author with Barnes and Mandel. Thus, it is important that the authors set out a proper, written record of the methods and findings, which could serve as an accessible document that others can reference in connection with this work.

3 Methods

3.1 Study Phasing

3.1.1 Phase I

As noted earlier, the present study proceeded in two phases. Barnes initiated the first phase prior to his partnership with Mandel. With the aid of John Hannigan, a senior IAS analyst who was not involved in the original drafting of the pertinent intelligence reports and had no reporting relationship to Barnes, Barnes undertook to categorize all judgments reported in intelligence reports produced by his (MEA) division. That effort focused on the 51 reports produced by nine different intelligence analysts over roughly a 20-month period from March 2005 to October 2006. That categorization activity yielded 1,231 judgments of which 649 (52%) were judged to be of a predictive nature (i.e., forecasts) and assigned a numeric probability. We discuss the coding rules for categorizing judgments as predictive or not in section 3.2 of this report. Fifty of the original 51 reports included at least one forecast, and all nine analysts made forecasts.

For each of the forecasts, outcome data was sought. Hannigan, the senior IAS analyst working with Barnes (hereafter Coder 1) coded the outcomes. The analysis of outcomes was sensitive to the event time frame specified in the forecast. Thus, coding was conducted at least several months after a forecast was given. Coder 1 initially assigned one of five outcome codes to each forecast: (a) the forecasted event occurred, (b) the forecasted event did not occur, (c) the forecasted event partially occurred, (d) the forecasted event partially did not occur, and (e) the outcome of the forecasted event could not be coded. However, for the quantitative analyses later conducted by Mandel and his team, the partial categories, with their inherently ambiguous status, were treated as unscorable cases as well. Of the 649 forecasts, 580 (89%) had outcomes that were coded as either having occurred or not having occurred. This subset constitutes the sample of forecasts from Phase I of the present study that are analyzed further. Finally, note that the terms "occurred" and "did not occur" are used in this study in a content-neutral manner. Thus, the event described in a judgment could be an omission, commission, positive or negative rate change, and so on. For instance, if an analyst forecasted that it was *extremely likely* [9/10 chance] that *X* would not happen, and *X* in fact did not happen, then this was coded as an occurrence. That is, the event in this case is "*X* does not occur" and not simply "*X*." In many cases, a forecast about *X* occurring could have been reframed as a statement about *X* not occurring or vice versa (e.g., "*extremely unlikely* [1/10 chance] that *X* will occur").

3.1.2. Phase II

By 2009, the analysis of forecasts from Phase I was completed, and Barnes and Mandel started to plan for a second phase of intelligence report coding and subsequent data analysis. Rex Brynen, a professor of political science at McGill University, was contracted as a second coder (hereafter Coder 2) and completed the coding of Phase II over the summer of 2010.

As in Phase I, Phase II focused on examining the quality of strategic intelligence forecasts. However, in addition to coding 73 MEA Division reports, Phase II broadened the examination to reports produced by other IAS divisions—including Asia (24 reports), Europe (10 reports), Global (2 reports), and Western Hemisphere (18 reports)—as well as the interdepartmental Intelligence Assessment Coordinating Committee (IACC) and Deputy Minister Intelligence Assessment Committee (DMIAC) (17 reports). In all, Phase II incorporated forecasts produced by 33 intelligence analysts.

3.1.3. Overview of Judgments Abstracted in Phases I and II

Table 1 provides an overview of the judgments extracted in Phases I and II. The first column indicates the study phase (I or II) and the organizational cluster (i.e., IAS division or interdepartmental committee). In the latter case, various partitions that may be of interest to the reader are reported. For instance, because the MEA Division is the only organizational cluster to use numeric probabilities as part of its standard operating procedure, Table 1 also summarizes the counts for all IAS MEA judgments (i.e., across Phases I and II), for non-MEA judgments (all of which were collected in Phase II), and for the subset of the latter that are IAS judgments. Table 1 also provides summary counts for Phase II over all organizational clusters and for the overall sample across Phases I and II.

Table 1: Numbers of Total Judgments, Total Forecasts, Determinate Forecasts, and Scorable Forecasts by Study Phase and Organizational Cluster

Phase / Organizational Cluster	Total Judgments	Total Forecasts	Determinate Forecasts	Scorable Forecasts
I/ IAS MEA	1231	649 (53%)	649 (100%)	580 (89%)
II/ IAS MEA	1299	714 (55%)	630 (88%)	493 (78%)
II / IAS Asia	244	144 (59%)	110 (76%)	86 (78%)
II / IAS Europe	120	66 (55%)	55 (83%)	46 (84%)
II / IAS Global	58	35 (60%)	20 (57%)	14 (70%)
II / IAS Western Hemisphere	205	105 (51%)	84 (80%)	69 (82%)
II / Interdepartmental	429	220 (51%)	173 (79%)	134 (77%)
All IAS MEA	2530	1363 (54%)	1279 (94%)	1073 (84%)
All Phase II	2355	1284 (55%)	1072 (83%)	842 (79%)
All non-MEA	1056	570 (54%)	442 (78%)	349 (79%)
All IAS non-MEA	627	350 (56%)	269 (77%)	215 (80%)
All Phases and Clusters	3586	1933 (54%)	1721 (89%)	1422 (83%)

The second column of Table 1 reports the total number of judgments extracted from the relevant cluster. Because not all judgments are forecasts, the third column reports the specific number of forecasts that were extracted from the relevant cluster and the percentage of the total judgments that the forecasts comprised. The fourth column shows the number of forecasts that were assigned numeric probabilities and the percentage of the total number of forecasts that this subset comprised. Note that in Phase I, all forecasts were assigned numeric probabilities by the MEA analysts who issued the forecast. Thus, 100% are determinate. In Phase II, in comparison, MEA analysts were instructed not to provide numeric probabilities for forecasts that used indeterminate verbal expressions of uncertainty. The specific expressions noted were “could,” “possible,” “might,” “may,” “a chance,” and “has the potential to.” Use of these terms was not forbidden in the MEA Division, but it was discouraged. Likewise, for Phase II forecasts from outside the MEA Division, Coder 2 did not assign numeric probabilities to the same set of indeterminate terms. The varying procedure from Phase I to Phase II is important to bear in mind when considering the values presented in Columns 3 and 4 of Table 1.

The final column in Table 1 reports the number of scorable determinate forecasts coded with a clear “occurred” or “did not occur” outcome. The values in this column exclude the partial and uncoded cases noted in 3.1.1. The percentages reported in the final column are the number of scorable forecasts over the number of determinate forecasts.

Table 2 provides a breakdown of the remainder of unscorable forecasts that were not coded into clear occurrence or non-occurrence categories. As can be seen, overall, 81% received a "partial" code, whereas the remaining 19% were coded as "unresolved" at the time of coding. The 56 cases in this latter category were assigned to one of three subsets of cases. The majority (n = 30) were deemed to be unfalsifiable forecasts. That is, for predictive judgments, the coder also assessed whether the judgment was worded in such a way that an observer could readily determine whether the event had occurred within the time period indicated. The judgment was coded as falsifiable if the outcome it described was clearly defined as a unique and observable event, on which a determination could clearly be made whether it had taken place during the time period of the assessment. The judgment was coded as unfalsifiable if it did not meet this requirement. The coder considered whether the judgment was a tautology, that is, true no matter what the outcome (e.g., "Leader X will likely invade country A unless he is persuaded not to"). To be falsifiable, the event needed to be worded in such a way as to be reasonably measurable within the given time frame. For example, "X is unlikely to launch military strikes against Y in the next year" has a clear outcome and is reasonably measurable, whereas "the influence of country Q in the region will increase over the next year" is in many aspects a subjective opinion which can vary among observers and would be more difficult to measure. Judgments that were unfalsifiable could not be coded. The remainder of uncoded judgments (n = 26) were either cases in which the outcome of the forecast had yet to be resolved or where the coder lacked necessary information to make a sound determination of the outcome.

Table 2: Breakdown of Unscorable Forecasts by Subtype, Study Phase, and Organizational Cluster

Phase / Organizational Cluster	Unscorable Forecasts	Partially Present	Partially Absent	Uncoded Forecasts
I/ IAS MEA	69	31 (45%)	26 (38%)	12 (17%)
II/ IAS MEA	137	63 (46%)	56 (41%)	18 (13%)
II / IAS Asia	24	13 (54%)	8 (33%)	3 (13%)
II / IAS Europe	9	3 (33%)	5 (56%)	1 (11%)
II / IAS Global	6	2 (33%)	3 (50%)	1 (17%)
II / IAS Western Hemisphere	15	6 (40%)	6 (40%)	3 (20%)
II / Interdepartmental	39	13 (33%)	8 (21%)	18 (46%)
All IAS MEA	206	94 (46%)	82 (40%)	30 (15%)
All Phase II	230	100 (43%)	86 (37%)	44 (19%)
All non-MEA	93	37 (40%)	30 (32%)	26 (28%)
All IAS non-MEA	54	24 (44%)	22 (41%)	8 (15%)
All Phases and Clusters	299	131 (44%)	112 (37%)	56 (19%)

3.2 Coding and Phase Harmonization Procedures

Phase II implemented a number of important changes from Phase I. As already noted, whereas Phase I was restricted to MEA Division judgments, Phase II included other IAS divisions as well as judgments from interdepartmental committees.

3.2.1. Judgment Categorization

As noted earlier, an effort was undertaken to categorize all judgments that appeared in the intelligence reports examined in this study. That effort had the same functional aim across the two phases—namely, to isolate forecasts from the total set of judgments conveyed in the reports.

However, the methods for doing so varied somewhat across the phases. This was mainly due to the fact that Phase II included reports from divisions other than MEA within IAS as well as from interdepartmental committees. Not only do the non-MEA Division reports not include numeric probabilities for forecasts, the analysts that produced those reports also were not required to categorize their judgments as predictive or otherwise. Thus, it is helpful to contrast the procedure for categorizing judgments in the MEA and non-MEA Divisions, respectively.

3.2.1.1. Judgments from the MEA Division

Within the MEA Division, analysts provided the first pass at categorizing their judgments as predictive or otherwise. In most cases, where judgments were of a predictive nature, analysts assigned a numeric probability value according to the divisional standard described in Section 3.2.2. However, as noted in Section 3.1.3, a proportion of predictive judgments (detailed earlier in Table 1) were issued with vague terms, such as "possible" or "might," which were not part of the lexicon described in that standard. In Phase I, these "indeterminate probabilities" were still given a numeric equivalent by analysts. In such cases, the numeric probabilities assigned were in the 40% to 60% range. By the time the judgments in Phase II were made, however, analysts were no longer permitted to assign numeric probabilities to such terms, and they were instead marked as "X/10." As for the majority of forecasts, which were issued with numeric probabilities, most were expressed as unconditional forecasts (e.g., "it is highly likely [9/10] that *p* will happen in the next month"), whereas a small proportion was expressed conditionally (e.g., "if *p* happens in the next month, then it is highly likely [9/10] that *q* will follow within the year"). Unconditional forecasts were included in the study. Conditional forecasts were included only if the suppositional condition (i.e., the "if" statement) turned out to be true (in which case they were simply treated as unconditional forecasts).

In Phase I, the categorization of the judgments into predictive (i.e., forecast) or otherwise underwent a second pass in which Barnes and Coder 1 reviewed each judgment. In some cases, the coder deemed a judgment categorized by the analyst as predictive to be non-predictive. For instance, although a degree of uncertainty may have been assigned, the judgment might have been of an explanatory or speculative nature (e.g., "it is likely that *x* was the cause of event *y*"). In such cases, the coder recoded the judgment as non-predictive. Conversely, some judgments categorized by analysts as non-predictive were re-categorized as predictive. In such cases, the numeric term corresponding to the verbal expression of uncertainty that appeared in the judgment was entered as the value for that judgment. The procedure for mapping between verbal and numeric expressions of uncertainty is described in Section 3.2.1. In Phase II, the procedure was the same, except that Coder 2 conducted the second pass mainly on his own.

3.2.1.2. Judgments from non-MEA divisional sources

For non-MEA divisional sources collected in Phase II, Coder 2 [alone] categorized the judgments as predictive or otherwise. Probability terms that were coded as indeterminate "X/10" cases in the MEA Division in Phase II were coded likewise if they appeared in non-MEA forecasts. Thus, the relative frequency of such terms across organizational units can be gauged in Phase II.

3.2.2. Numeric Probability Scale

Even within the MEA Division subsample, there was one particularly important difference across study phases, which concerned the divisional procedure used to elicit numeric probabilities for forecasts. In Phase I, an 11-point numeric scale was used that increased in whole integers from 0 to 10 (out of 10). However, prior to the start of Phase II, Barnes changed the scale to a 9-point version in which the values of 2 and 3 and, likewise, 7 and 8 were combined. The rationale for

this procedural change was that analysts in his division seemed to have a particularly difficult time distinguishing between these adjacent values, and that keeping them separate unnecessarily increased the number of levels of probability that analysts could realistically distinguish.

Thus, in Phase I, MEA Division judgments were based on an 11-point numeric scale, whereas in Phase II, they were based on a 9-point numeric scale. In order to harmonize data for quantitative analyses of judgment quality, Phase I judgments were recoded onto the 9-point scale. For data-analytic purposes, the "2 or 3" probability level was recoded as 2.5 (out of 10) and, likewise, the "7 or 8" probability level was recoded as 7.5 (out of 10).

It is important to note that, although MEA Division analysts were instructed to assess the likelihood of the forecasted events in numeric terms, the final reports used only verbal expressions of uncertainty, not the numbers themselves. Table 3 shows the mapping standard used in Phase II, once the 9-point scale was implemented.

The mapping standard was developed by the IAS MEA Division based on a review of the intelligence and behavioural psychology literature dealing with how verbal probability terms are commonly understood by readers. This was then adapted for the practical needs of an analytic group, particularly the requirement to have sufficient synonyms at each level of probability to allow for stylistic flexibility in presenting the analysis.

Note, too, that even though MEA Division analysts were advised to think in terms of the numeric probabilities they wanted to assign, there was in fact no way to ensure that they did so. Realistically, in many cases in which subjective probabilities were assigned, the first pass by the analyst in formulating a judgment may have been of a fuzzier, verbal nature. As well, as noted earlier, some proportion of forecasts were made using uncertainty terms, such as "it is possible that…" or "*x* might happen," which were not part of the lexicon and which were not assigned a numeric equivalent.

Although most of the forecasts were expressed in absolute terms, some were made in relative terms (e.g., "scenario *a* is the most likely, followed by scenario *b*"). In such cases, the analyst would sometimes assign a numeric probability value to the alternative events. Thus, in the preceding example, scenario *a* might be given a 6/10 and scenario *b* a 4/10 score. In other cases, only the rank order of the probabilities assigned to various outcomes was specified. In these cases, numeric probabilities were inferred by the coder based on the context in which the ranked alternatives appeared.

Finally, it is worth clarifying at the outset that the forecasts analyzed in this study are not, strictly speaking, *analysts'* forecasts, although they are analytic forecasts. Analysts are first and foremost accountable to their divisional directors, who review and challenge their assessments before the final draft is released to stakeholders. The drafting process also normally involves extensive consultation with other experts and government officials. Thus, the forecasts that appear in the intelligence reports are seldom the result of a single individual arriving at his or her judgment. It is more accurate to regard an analytic forecast as an organizational product reflecting the input of the primary analyst, the analyst's director, and possibly a number of peer analysts.

Table 3: IAS MEA Division Mapping Standard for Reporting Verbal Expressions of Uncertainty

Verbal Expression	Probability	Remark
will is certain	[10/10]	Where you can envisage no plausible scenario—however remote—where this event would not happen.
almost certain extremely likely highly likely	[9/10]	There remains some conceivable scenario—albeit very remote—that this event would not happen.
likely probable, probably	[7-8/10]	
slightly greater than even chance	[6/10]	Use rarely, only when there is a specific reason to judge the probability at greater than even but cannot be categorized as "likely."
Even chance	[5/10]	
slightly less than even chance	[4/10]	Use rarely, only when there is a specific reason to judge the probability at less than even but cannot be categorized as "unlikely."
unlikely (only a) low probability probably not	[2-3/10]	
very unlikely highly unlikely extremely unlikely little prospect	[1/10]	There remains some conceivable scenario—albeit very remote—that this event could happen.
no prospect will not	[0/10]	Where you can envisage no plausible scenario—however remote—where this event could happen.

3.2.3. Inferred Numeric Probabilities in the Non-MEA Division Subsample

The order of mappings (numeric to verbal or vice versa) is of significant importance to the current aim of tracking forecast quality using quantitative measures of calibration and discrimination. In the MEA Division subsample, the data provided by analysts provide the basis for a true calibration study because the predictions made were in fact numeric. We can therefore assess the quality of those forecasts in objective terms regardless of the validity of the numeric-to-verbal mapping process.

In contrast, the forecasts made by analysts in reports produced by other IAS divisions or by interdepartmental committees were issued in verbal terms. For these forecasts, a numeric equivalent had to be inferred by Coder 2. That coding followed a semi-structured process. In cases in which a verbal term matched one of the verbal terms listed in Table 3, the numeric equivalent assigned in the standard was also assigned as the inferred probability of the forecasted event. In cases in which a verbal term did not match one of the terms listed in Table 3, Coder 2 judged which of the terms in the standard was closest in meaning to the expressed term, and then used the numeric equivalent from the standard as the inferred probability of the forecasted event.

Therefore, the non-MEA Division data provide the basis for a quantitative judgment analysis in only a qualified sense. The integrity of the analysis depends largely upon the reliability and

validity of the process by which the numeric probabilities were inferred. An obvious validity-compromising limitation of this endeavour is that we do not know whether the analysts who made the forecasts in verbal terms would agree with the numeric probabilities assigned to their forecasts. A perhaps less obvious, reliability-compromising limitation of the inference procedure is that we cannot be sure that other independent coders would have assigned verbal terms not in the MEA divisional standard to the same probability level. That is, they might have decided that a particular term's nearest neighbour was one that had a different numeric equivalent.

Accordingly, the analysis of forecasts from MEA Division and non-MEA Division sources must be interpreted with differing degrees of caution and understood for what they are, respectively: a study of bona fide numeric forecasts in the first instance, and a study of inferred numeric forecasts in the second instance. Both are of value, as is their comparison. However, one should also bear in mind their differences when interpreting the findings.

3.2.4. Inter-rater Reliability Analysis

In planning for the initiation of Phase II, it was decided that Coder 2 would independently code a small sample of reports from Phase I so that an estimate of inter-rater reliability could be established. Two reports, one on the Middle East and the other on Africa, were recoded. These two reports included 51 predictive judgments. Coder 2 blind coded the outcomes of these forecasted events. Using the five-category coding scheme (i.e., including partial outcomes), there were 46 exact hits, all of which were on unambiguous "occurred" or "did not occur" cases. Thus, on this conservative measure of agreement, there was a 90.2% agreement rate (i.e., 46/51). When recoded as a three-category outcome scheme (i.e., "occurred," "did not occur," or "undetermined"), that rate increased to 92.2% (47/51). Coder 2 attributed the few disagreements to a difference in substantive interpretation in two cases, a difference in semantic interpretation of a term in one case, and to new information becoming available that Coder 1 would not have had access to in one case.

It is reassuring to see that the inter-rater reliability, though not perfect, was very high, in spite of the fact that the events being judged were often complex in nature. This finding should serve to strengthen the reader's confidence that the coding of outcomes for forecasted events was not a mere exercise in subjectivity. Independent coders, unaware of each other's assessments, strongly agreed on what actually happened.

3.2.5. Potential Moderator Variables

In addition to the coding of outcomes, several other variables were coded. These variables serve as potential moderators of the judgment quality measures used in this study.

3.2.5.1. Analyst's experience

The second author, Barnes, coded the experience level of analysts into "junior" and "senior" categories. In some cases, an analyst who was junior in Phase I was senior in Phase II and was coded as such. In 38 cases of IAS predictions, multiple analysts were associated with a judgment, and in those cases a "multiple" code was assigned, signifying that they should be treated as missing data for analyses involving analyst experience. Those cases were all associated with MEA Division judgments. Moreover, every forecast made in an interdepartmental report was made by multiple analysts. Accordingly, those forecasts were not coded in terms of experience.

3.2.5.2. Forecast difficulty

The coder coded the difficulty of the forecast as low-to-moderate or moderate-to-high. In Phase I, the coder made these assessments without input from the analyst, whereas in Phase II the analysts were asked to provide an initial "easy" or "hard" coding. For simplicity, we refer to these levels of difficulty later in the report as "easier" and "harder," respectively. Forecasts of low/moderate difficulty were defined as judgments under most or all of the following conditions:

(1) availability of a substantial and credible information base,

(2) involving a limited number of factors and/or largely a straight-line continuation of current trends,

(3) little influence of irrational or unpredictable behaviour, or

(4) generally involving a short time horizon (several months).

Forecasts of moderate/high difficulty were defined as judgments affected by some of the following conditions:

(1) a limited and unreliable information base,

(2) involving a wide range of complicated factors with multiple potential outcomes,

(3) high likelihood of unpredictable behaviour, or

(4) involving a longer time horizon (a year or more).

3.2.5.3. Forecast importance

The importance of the forecast for intelligence consumers was coded as "low/moderate" and "moderate/high" by the coder. Once again, for simplicity, these levels are referred to later in the report as "lower" or "higher" levels of importance. Forecasts of low/moderate importance were defined thus:

> The event being considered has negligible or limited impact on specific Canadian interests and has negligible or limited impact on the overall stability of the country/region in question or on international diplomacy, the global economy, or other important international issues (proliferation, illegal migration, etc.).

Forecasts of moderate/high importance were defined thus:

> The event being considered has either (1) substantial or very significant impact on Canadian diplomatic, security, economic, or consular interests; or (2) substantial or very significant impact on the overall stability of the country/region in question or on international diplomacy, the global economy, or other important international issues (proliferation, illegal migration, etc.).

3.2.5.4. Forecast prominence

The prominence of the forecast in the report was coded based on whether it was a key judgment or not. Key judgments are those identified at the outset of the report in what is akin to an executive summary. They may overlap with judgments expressed later in the main body of the report (non-key judgments), and in such cases only the key judgment was recorded so that the same forecast was not entered twice.

3.2.5.5. Report origin

Finally, in Phase II only, Barnes recorded whether a forecast was from a report requested by an intelligence consumer or whether it was from a report resulting from an internally generated tasking.

3.2.5.6. Region of interest

For the MEA Division sample, forecasts about the Middle East were compared with forecasts about Africa.

3.2.5.7. Forecast source

For the analysis of inferred probabilities (i.e., from sources other than the MEA Division), a comparison was made between forecasts generated by the IAS and those generated by interdepartmental committees.

3.2.5.8. Forecast time frame

In Phase II, time frame information was extracted from reports and from judgments. In the former case, a report might set an approximate time frame over which its assessments were intended to apply. In some cases, individual judgments also provided a time frame. For instance, the report might specify a six-month time frame overall, yet a specific forecast in that report might specify a longer (e.g., one year) or shorter (e.g., less than one month) time frame. In cases where time frame information was presented for the overall report and for a specific forecast, the time frame pertaining to the forecast was used as the basis for coding that forecast. In cases where no explicit information about time frame was given, the coder for this variable (Barnes) inferred the time frame from the context of the report and judgment. Five coding categories were used, which corresponded to the most common time frame ranges covered by reports. These were (a) up to one month, (b) two to three months, (c) three to six months, (d) six months to one year, and (e) more than one year.

3.3 Data Transfer and Scope of Accessibility

The intelligence reports from which the forecasts studied here were drawn are classified documents requiring Top Secret/Special Access clearance for viewing. In order to facilitate the analyses conducted in this study, the judgments being assessed were separated from their content, thus enabling their processing at the unclassified level. The pertinent data was transferred from an IAS database to DRDC Toronto for analysis for the purpose of this study. With the exception of the first author (Mandel), researchers at DRDC Toronto did not read any of the classified reports from which the forecasts were drawn.

3.4 Data-analytic Procedures

This study uses several standard, quantitative measures of judgment quality in order to shed light on the forecasting performance of intelligence analysts. We report a range of measures, some of which are easy to grasp (e.g., the percentage of correct forecasts, where "correct" means being on the appropriate side of "fifty-fifty"), and others of which may require some explanation.

3.4.1. Measures of the Percentage of Correct Forecasts

At the "easy" end of the spectrum, we calculate the percentage of correct forecasts in two ways. First, ignoring forecasts that are "right on the fence"—namely, fives on the 0-10 probability scale used in this study—we code as "correct" those forecasts that were issued with (a) probabilities of six or greater, where the forecasted event did in fact occur; and (b) probabilities of four or less, where the forecasted event did not in fact occur. The sum of these values, divided by the total number of forecasts (excluding any fives) is our *liberal* measure of percentage correct ("percentage of correct forecast by liberal criterion"—"$\%CF_L$"). The measure is liberal in the sense that it does not penalize analysts for making fence-sitting forecasts.

We also calculate a *conservative* measure of percentage correct ("percentage of correct forecast by conservative criterion"—"$\%CF_C$"), which adds any fives into the frequency of the denominator, thus counting all "fence-sitting" forecasts as wrong. From these two measures of percentage correct, one might in fact choose to consider a *moderate* measure based on their average. Such a measure would punish fence-sitting behaviour but only half as much as being on the wrong side of the fence. Given the ease of computing such a measure from the other two, we report only the liberal and conservative measures. In practice, fence-sitting judgments were rare in this sample, and, thus, the two measures tend to be close in value.

Although percentage correct measures are easy to understand, they provide only rough measures of forecast quality. They are rough in the sense that correctness and incorrectness are insensitive to variations in probability level on either side of the fence (that is, left or right of fifty-fifty). For instance, forecasting an event that actually occurs would be treated as no more correct if issued with a 90% probability than a 60% probability. Thus, while the liberal measure of percentage correct ignores fence sitting and the conservative measure punishes it, both measures fail to distinguish between forecasts that sit "near" versus "far" from the fence.

3.4.2. The Probability Score and its Partitions

The remaining measures we report, although potentially harder to intuitively grasp, present a more detailed picture of analysts' forecast quality. The measures we pay closest attention to are *calibration* and *discrimination*. Both of these measures are components (or in the case of the discrimination measure used here, an adjusted component) of the Brier Score, also known as the Probability Score (*PS*) (Brier, 1950; Murphy, 1973).

As Yaniv et al. (1991) point out, *PS* is decomposable into components that are more readily interpretable as measures of judgment quality. Thus, while we report *PS*, our analytic focus is on calibration and discrimination.

To begin, assume a sample of N cases where the outcome of each is coded dichotomously such that $d = 1$ if the outcome is present and $d = 0$ if the outcome is absent. Assume further that, for each case, a probability of the outcome, f, is assigned. The resulting series of judgment-outcome pairs can be shown as a two-way table (Table 4) for the 9-point scale used in the present study.

Table 4: Case Classification by Probability Judgment Category and Outcome

Outcome (*d*)	**Probability judgment category (f_j)**									
	0%	10%	25%	40%	50%	60%	75%	90%	100%	Total
Present ($d = 1$)	$N_{1,A}$	$N_{1,B}$	$N_{1,C}$	$N_{1,D}$	$N_{1,E}$	$N_{1,F}$	$N_{1,G}$	$N_{1,H}$	$N_{1,I}$	N_1

Absent ($d = 0$)	$N_{0,\,A}$	$N_{0,\,B}$	$N_{0,\,C}$	$N_{0,\,D}$	$N_{0,\,E}$	$N_{0,\,F}$	$N_{0,\,G}$	$N_{0,\,H}$	$N_{0,\,I}$	N_0
Total	N_A	N_B	N_C	N_D	N_E	N_F	N_G	N_H	N_I	N

As Table 4 shows, cases may be partitioned by the assigned probability judgment category and by the outcome of the case. As Yaniv et al. (1991) put it, "In a sense, the judge sorts the events by their probabilities (horizontally) and 'nature' sorts the events by their outcomes (vertically)" (p. 612).

Based on the notation in Table 4, we define the relative frequency of the outcome in each probability judgment category as

$$\bar{d}_j = \frac{N_{1,j}}{N_j}.$$

The base rate with which the outcomes occur is defined as

$$\bar{d} = \frac{N_1}{N}.$$

PS, which provides a global measure of accuracy across N cases, is calculated as follows:

$$PS = \frac{1}{N}\sum_{n=1}^{N}(f_n - d_n)^2,$$

where f_n denotes the nth probability judgment and d_n denotes the nth outcome.

PS is a linear combination of three sums of squares, which represent measures of outcome variance, discrimination, and calibration, respectively, and can be expanded as follows:

$$PS = \bar{d}(1-\bar{d}) - \frac{1}{N}\sum_{j=1}^{J} N_j(\bar{d}_j - \bar{d})^2 + \frac{1}{N}\sum_{j=1}^{J} N_j(f_j - \bar{d}_j)^2.$$

Thus, one can alternatively express *PS* as a linear combination of three partitions:

$$PS = \text{VI} - \text{DI} + \text{CI},$$

where VI indexes the variance of the outcome, DI is the discrimination index, and CI is the calibration index.

3.4.2.1. The calibration index

In the present study, CI constitutes our measure of calibration. An analyst is better calibrated to the extent that his or her probability judgments match the relative frequencies of outcomes occurring within the relevant judgment category. Figure 1, which shows a hypothetical *reliability*

diagram (Yates, 1990), shows a case of perfect calibration, where the relative frequency of outcome occurrence in each probability judgment category is equal to the value of that category.

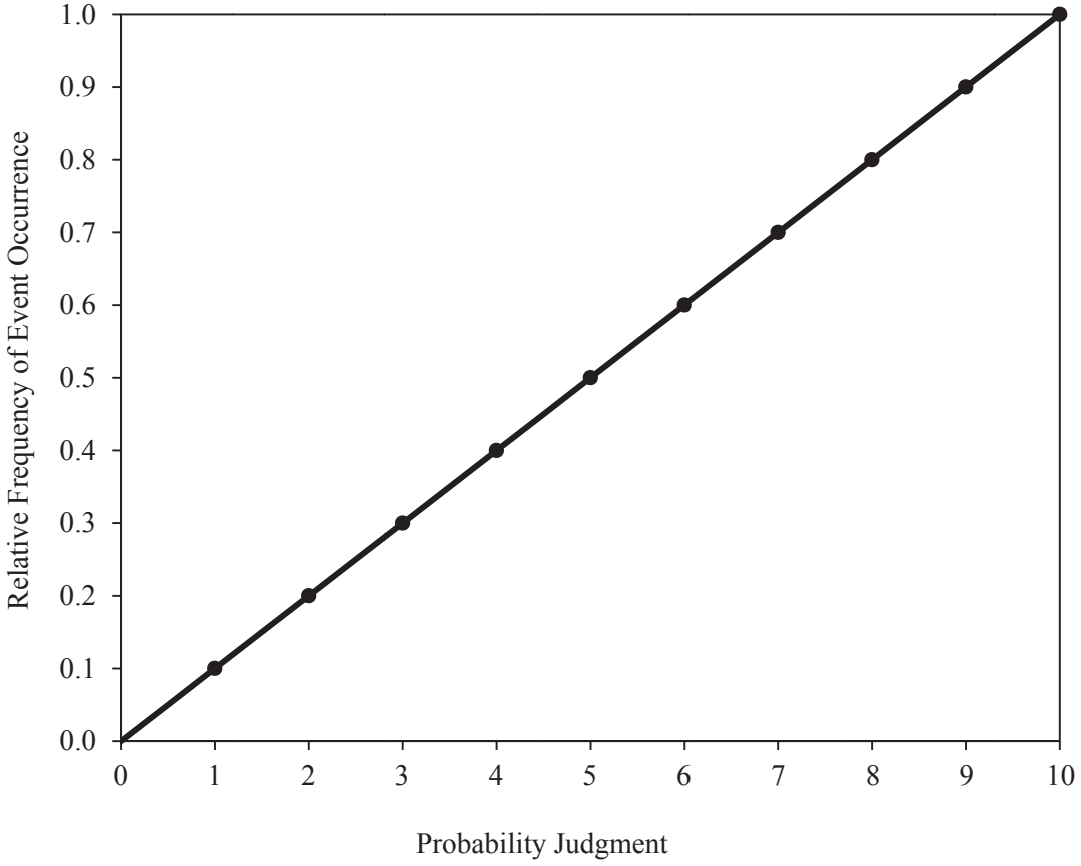

Figure 1: Calibration Curve Showing Perfect Calibration

Deviations from the 45° perfect calibration line show the degree of miscalibration within a given judgment category. Because the number of cases may vary across judgment categories, the overall degree of calibration may be difficult to read off of the "calibration curve" plotted in a reliability diagram. In the present report, we plot data in reliability diagrams for all judgment categories, but we only connect the data points forming the calibration curve for those categories that have 10 or more cases. We do so in order to minimize the perceptual salience of the most unreliable regions of the curve as well as to smooth the curves somewhat.

In addition to the reliability diagram, we report the calibration index (CI), which provides a summary measure across all judgment categories and cases. It is scaled from 0 to 1, with 0 representing perfect calibration. In practice, it is useful to examine both CI and the reliability diagram since the latter often reveals noteworthy characteristics about the quality of forecasts. In particular, the shape of the calibration curve will tend to indicate whether the source of miscalibration is mainly due to overconfidence or underconfidence.

Figure 2, which is adapted from Koehler, Brenner, and Griffin (2002), shows some characteristic patterns of miscalibration. Overprediction bias occurs when all (or, less stringently, most)

judgments are higher than the corresponding outcome relative frequency. Conversely, underprediction bias occurs when all or most judgments are lower than the corresponding outcome relative frequency. Overextremity bias occurs when judgments are too extreme, such that judgments below 50% are too low and those above 50% are too high. In the present research, overextremity bias reflects overconfidence in one's hypothesis because the certainty an analyst assigns to events occurring or not occurring is too great. Conversely, underextremity bias occurs when judgments are too conservative, such that those below 50% are too high and those above 50% are too low. In the present research, underextremity bias reflects underconfidence in one's hypothesis because the certainty an analyst assigns to events occurring or not occurring is not enough.

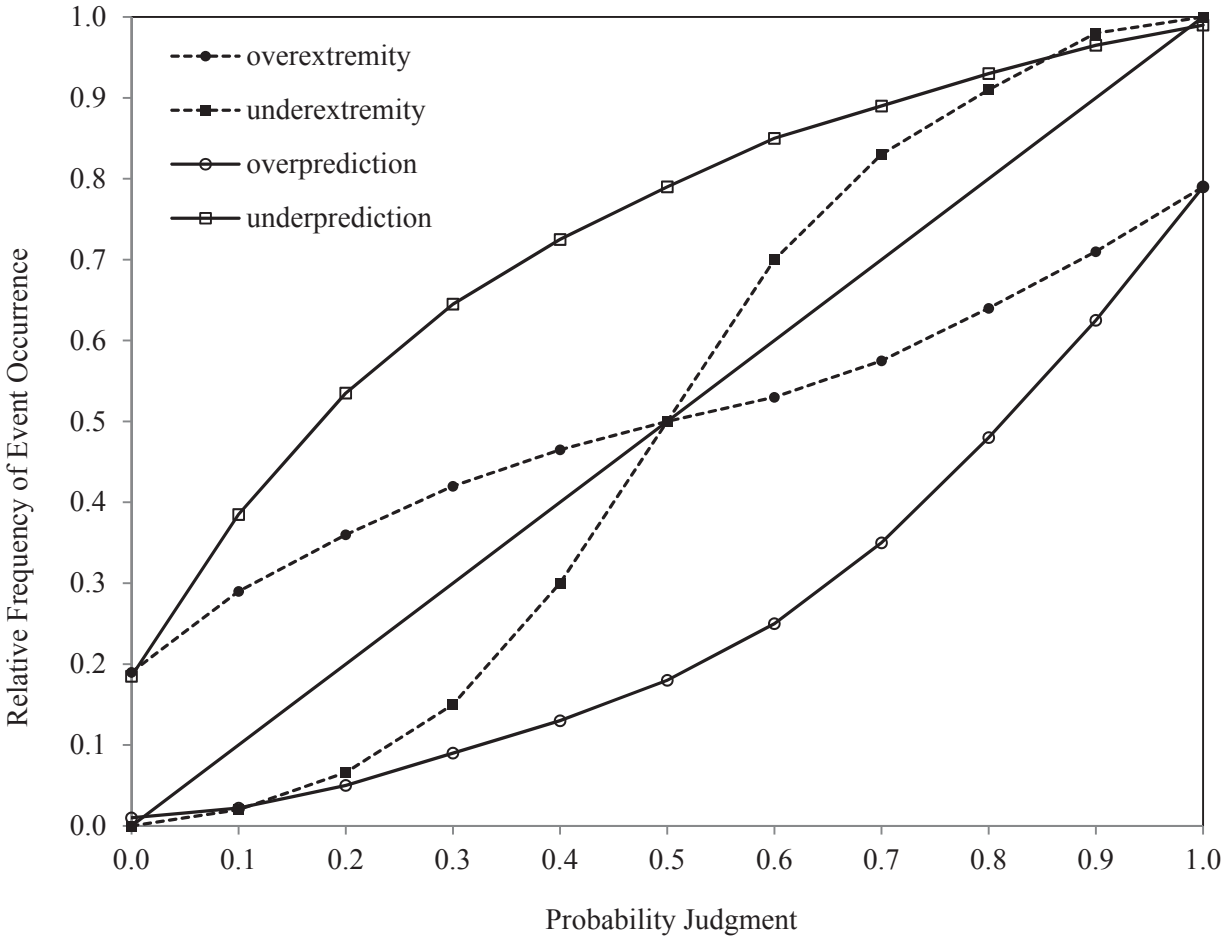

Figure 2: Four Characteristic Patterns of Miscalibration

3.4.2.1. The discrimination index and adjusted eta-squared

In addition to calibration, we also report a corrected measure of discrimination, the second component of *PS*. Discrimination measures the degree to which one can differentiate between cases in which an event will occur from cases in which it will not occur. Discrimination has also been called *resolution* in some literature (Murphy, 1973) since it measures the degree to which one can resolve cases into their correct classification.

The discrimination index (DI) is a weighted average of the distance of the relative frequency of event occurrences within each judgment category ($\bar{d}_j$) from the base rate of event occurrence across all categories ($\bar{d}$). Thus, higher values of DI indicate better discrimination, and the poorest discrimination would occur if all predictions were identical to the base rate ($\bar{d}$). It is noteworthy that, in this extreme case of poor discrimination, calibration would be perfect. The example illustrates a more general principle that good calibration does not imply good discrimination, and, likewise, good discrimination does not imply good calibration.

In the present study, we adopt the amendments to DI proposed by Yaniv et al. (1991). The first of those normalizes DI (also proposed by Sharp, Cutler, & Penrod, 1988), yielding the normalized discrimination index η^2:

$$\eta^2 = \frac{\mathrm{DI}}{\mathrm{VI}}.$$

The normalization of DI is useful because DI is in fact a component of VI, called *variance between*, which represents the variances accounted for by the judgment categories $f_1, f_2, \ldots, f_J$. The second component of VI, called *variance within*, represents the variance not accounted for by the judgment categories. Thus, η^2 is a measure of the proportion of variance explained by the judgment categories over the total variance. For that reason, it is equivalent to η^2 in a one-way analysis of variance, making the measure readily interpretable as the proportion of outcome variance explained by the forecasts.

The second amendment corrects for bias produced by variations in the ratio of the number of judgment categories J to the number of cases N. This bias follows from the fact that the expectation of η^2 under nil discrimination is positive:

$$\eta_0^2 = \frac{J-1}{N}.$$

The adjusted value of η^2 — i.e., η^{2*} — corrects for this bias such that the measure equals 0 when discrimination is nil, even under a high category-to-case ratio. The correction is calculated as follows:

$$\eta^{2*} = \frac{1-\eta^2}{1-\eta_0^2} = \frac{N \bullet \eta^2 - J + 1}{N - J + 1}.$$

3.4.2.3. Adjusting the probability score to control for forecasting difficulty

Improvements in the probability score *PS* can be brought about by improving forecaster skill or by reducing the unpredictability of the forecasting environment. The greater the unpredictability of the environment, the more difficult the task of forecasting will be. Recall that *PS* is partitioned into three components: CI, DI, and VI. The first two, as already noted, measure aspects of forecaster skill, whereas VI measures the unpredictability of the forecasting environment. For binary outcomes, as in the present study, VI is greatest when the base rate of event occurrence is .50, in which case VI = .25. This corresponds to the case of maximum forecasting difficulty. Accordingly, a forecaster *A* with *PS* = .10 would demonstrate more skill if the base rate was .50 (VI = .25) than another forecaster *B* with the same *PS* value forecasting in an environment where the base rate was .80 (VI = .16).

In order to better capture the skill of a forecaster or a set of forecasts from different forecasters (as in the present study), we apply the adjustment proposed by Tetlock (2005), which is similar to an earlier proposal by Winkler (1994):

$$PS^* = \frac{PS_{\bar{d}} - PS}{T}, \text{ where } T = (1-\bar{d})^2 \text{ iff } f_n \geq \bar{d}, \text{ else } T = \bar{d}^2.$$

PS^* is the difficulty-adjusted probability score. $PS_{\bar{d}}$ is the probability score that would be obtained if the forecaster always predicted the base rate, $\bar{d}$. *PS* is the unadjusted probability score defined earlier. The denominator *T* depends on whether the forecaster's judgment (f_n) is equal or greater than the base rate or less than the base rate.

For binary outcomes, PS^* is computed as follows for the *n*th forecast:

$$PS_n^* = \frac{(\bar{d}-1)^2 - (f_n-1)^2}{(\bar{d}-1)^2} \text{ if the outcome occurred and } f_n \geq \bar{d} \text{ (hit),}$$

$$PS_n^* = \frac{(\bar{d}-1)^2 - (f_n-1)^2}{\bar{d}^2} \text{ if the outcome occurred and } f_n < \bar{d} \text{ (miss or underprediction),}$$

$$PS_n^* = \frac{\bar{d}^2 - f_n^2}{(\bar{d}-1)^2} \text{ if the outcome did not occur and } f_n \geq \bar{d} \text{ (false alarm or overprediction),}$$

$$PS_n^* = \frac{\bar{d}^2 - f_n^2}{\bar{d}^2} \text{ if the outcome did not occur and } f_n < \bar{d} \text{ (correct rejection).}$$

Difficulty-adjusted probability scores equal zero if the forecaster simply predicts the base rate, while positive values indicate skill above a simple base-rate prediction strategy and negative values indicate performance below the base rate. The maximum possible value of PS^* is 1 and it can be achieved at any base-rate level, while the minimum possible value is -3 when the base rate is .50 and approaches 1 as the base rate becomes more extreme (i.e., closer to 0 or 1).

To illustrate, imagine a forecaster who predicted event occurrence with 100% certainty when the base rate is .50 and the event actually occurred. In this case,

$$PS_n^* = \frac{(\bar{d}-1)^2 - (f_n-1)^2}{(\bar{d}-1)^2} \text{ given that the outcome occurred and } f_n \geq \bar{d} \text{ (hit). Specifically,}$$

$$PS_n^* = \frac{(.5-1)^2 - (1-1)^2}{(.5-1)^2} = 1.$$

Now, imagine instead that the event, in fact, did not occur. In this case,

$$PS_n^* = \frac{\bar{d}^2 - f_n^2}{(\bar{d}-1)^2} \text{ given that the outcome did not occur and } f_n \geq \bar{d} \text{ (false alarm or overprediction). Specifically,}$$

$$PS_n^* = \frac{.5^2 - 1^2}{(.5-1)^2} = \frac{-.75}{.25} = -3.$$

Finally, had the base rate been .10 instead of .50 in the last false-alarm example, then

$$PS_n^* = \frac{.1^2 - 1^2}{(.1-1)^2} = \frac{-.99}{.81} = -1.22.$$

4 Results

The results of this study are divided into three main sections. In Section 4.1, findings from the MEA Division are summarized. In Section 4.2, the findings based on the inferred probabilities drawn from non-MEA forecasts are presented. Finally, in Section 4.3, a combined analysis of the forecasts from all subsamples is shown.

Each of the three main sections presents a summary table that shows the sample size (*N*), the base rate of event occurrence ($\bar{d}$), the liberal and conservative percentage of correct forecasts (%CF_L and %CF_C, respectively), the probability score (*PS*), the calibration index (CI), the discrimination index (DI), the variance of the outcome (VI), adjusted eta-squared (η^{2*}), and the difficulty-adjusted probability score (PS^*). These statistics are presented for the overall sample summarized in each section as well as for each relevant subsample corresponding to the levels of the various factors described earlier. Following the summary table, a series of reliability diagrams showing the calibration curves for the relevant (sub)samples are also presented. Note that, although values based on less than 10 forecasts are shown as points in these diagrams, they are left disconnected from the calibration line due to their low reliability. The value of 10 is admittedly arbitrary. Readers are encouraged to pay attention to the actual frequencies of forecasts within each probability judgment category, which are displayed below the *x*-axis. As well, readers are encouraged to consult the statistics shown in the relevant table, which take the varying number of cases in each judgment category into account.

4.1 Middle East and Africa Division

As noted earlier, the MEA Division is treated separately because only in this subsample were analysts specifically instructed to assign numeric probabilities on a 0 to 10 integer scale to their forecasts. As Table 5 shows, 57% of the 1,073 forecasts isolated for this analysis involved events that actually occurred. Overall, forecast quality was very good. Roughly 89% of judgments were correctly classified when fence-sitting 5/10 judgments were counted as wrong, and that figure rose to 92% when the few 5/10s were omitted from the calculation. The calibration index value of 0.014 is impressive, especially in comparison to other studies of expert socio-political judgment (e.g., Tetlock, 2005). In terms of discrimination, η^{2*} is equally impressive, showing that analysts' forecasts are successfully accounting for about 68% of total outcome variance.

The reliability diagram for the overall MEA Division sample is shown in Figure 3. The overall shape of the calibration curve is consistent with an underextremity bias, except for the endpoints of 0/10 and 10/10. This deviation from the idealized underextremity curve shown in Figure 2 is not surprising because, at the extremes, miscalibration can only take the form of "overconfidence"—namely, being too extreme in that some of the extreme forecasts do not pan out. Moreover, the deviations are slight. About 94% of the 243 10/10 predictions actually occurred and about 93% of the 41 0/10 predictions did not occur. Thus, at the extremes, where forecasts would be most informative for intelligence consumers in general and decision makers in particular, accuracy was quite high.

Table 5: Summary Statistics for the MEA Division

	N	$\bar{d}$	%CP_L	%CP_C	*PS*	VI	CI	DI	η^{2*}	PS^*
Overall	1073	.57	91.8	89.0	.091	.245	.014	.168	.681	.618
Phase:										
I	580	.61	89.5	85.0	.110	.237	.013	.140	.583	.518
II	493	.52	94.4	93.7	.069	.250	.017	.197	.786	.613
Analyst experience:										
Junior	418	.63	89.4	86.8	.112	.234	.018	.139	.588	.495
Senior	617	.54	94.0	91.2	.074	.248	.016	.189	.760	.696
Multiple	38	.32	82.9	76.3	.133	.216	.043	.126	.473	.510
Difficulty:										
Easier	381	.73	95.8	95.5	.052	.198	.016	.162	.811	.661
Harder	609	.49	88.8	87.5	.113	.250	.016	.153	.609	.551
Importance:										
Lower	287	.77	93.1	89.9	.083	.179	.020	.116	.637	.427
Higher	786	.50	91.3	88.7	.094	.250	.014	.170	.675	.623
Prominence:										
Key	343	.55	92.5	89.8	.086	.248	.013	.175	.699	.651
Non-key	730	.58	91.5	88.6	.094	.244	.015	.165	.672	.593
Origin*:										
External	149	.50	95.8	95.8	.062	.250	.019	.207	.819	.753
Internal	344	.52	94.3	93.7	.072	.250	.017	.194	.771	.708
Region:										
Middle East	459	.56	91.4	88.2	.095	.247	.014	.165	.664	.603
Africa	614	.58	92.1	89.6	.088	.244	.014	.170	.693	.630
Time Frame*:										
Up to 1 month	19	.58	89.5	89.5	.107	.244	.037	.174	.503	.523
2 - 3 months	33	.55	93.9	93.9	.067	.248	.013	.194	.714	.733
3 - 6 months	178	.46	93.3	93.3	.075	.248	.013	.186	.739	.700
6 months - 1 year	246	.56	95.5	93.9	.065	.246	.021	.202	.814	.727
> 1 year	17	.29	100.0	100.0	.026	.208	.026	.208	1.00	.810

* Phase II data only.

The more general pattern of underextremity bias revealed in Figure 3 indicates that analysts are overly conservative in their forecasts. Most often, their judgments are on the correct side of fifty-fifty, but, for a significant subset of their forecasts, they communicate greater uncertainty than is warranted. Take, for instance, forecasts offered with a 20% to 30% chance of occurrence (2.5 on the *x* axis in Figure 3). These forecasts indicate that the event probably will *not* occur, but the indication is too weak given that far more than 75% do not occur. Indeed, just over 95% of those 153 forecasts did not occur. The same conservative tendency can be seen on the probability scale above the fifty-fifty mark. For instance, for forecasts in the 70% to 80% chance of occurrence range, instead of the expected result of 75% of the forecasted events occurring, about 87% actually occurred (a value that is closer to a 9/10 on the relevant scale).

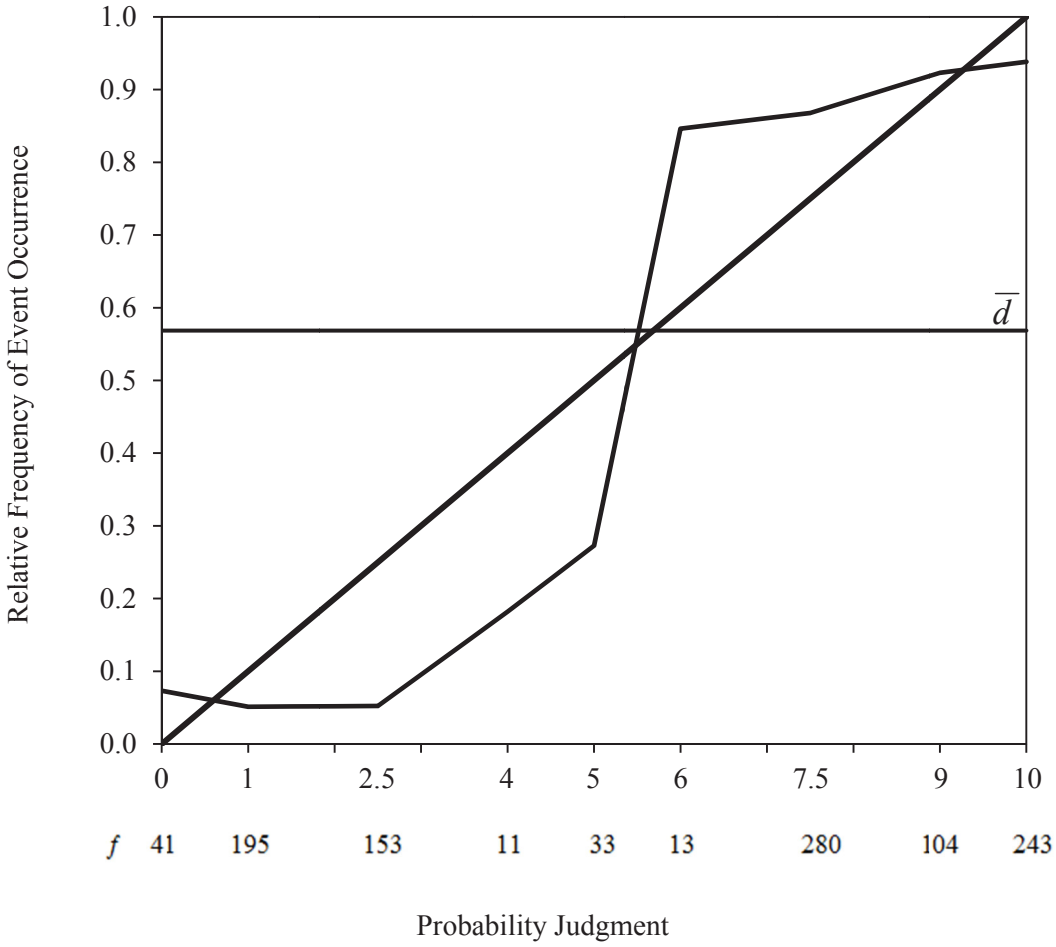

Figure 3: Reliability Diagram for the MEA Division Overall

In the following subsections, the findings from the MEA Division are analyzed in terms of the variables reported in Table 5.

4.1.1 Study Phase

The findings in Table 5 show that forecast quality was somewhat better in Phase II. Correct classifications were higher by about five percentage points using the liberal criterion and by about eight percentage points using the conservative criterion. Moreover, examining η^{2*}, one sees that approximately 79% of the total outcome variance was explained in Phase II as compared with 58% in Phase I. The performance advantage in Phase II is also evident in the difficulty-adjusted probability scores.

Figure 4 shows the calibration curves for Phases I and II. It is evident that correct classification at the extremes (i.e., 0/10 and 10/10) was better in Phase II (96% correct) than in Phase I (92% correct). Moreover, it does not appear that the differential treatment of indeterminate probabilities across study phases accounts for the differential level of performance. In Phase 1, as noted in section 3.2.1.1, such "probabilities" were all assigned mid-range numeric values.

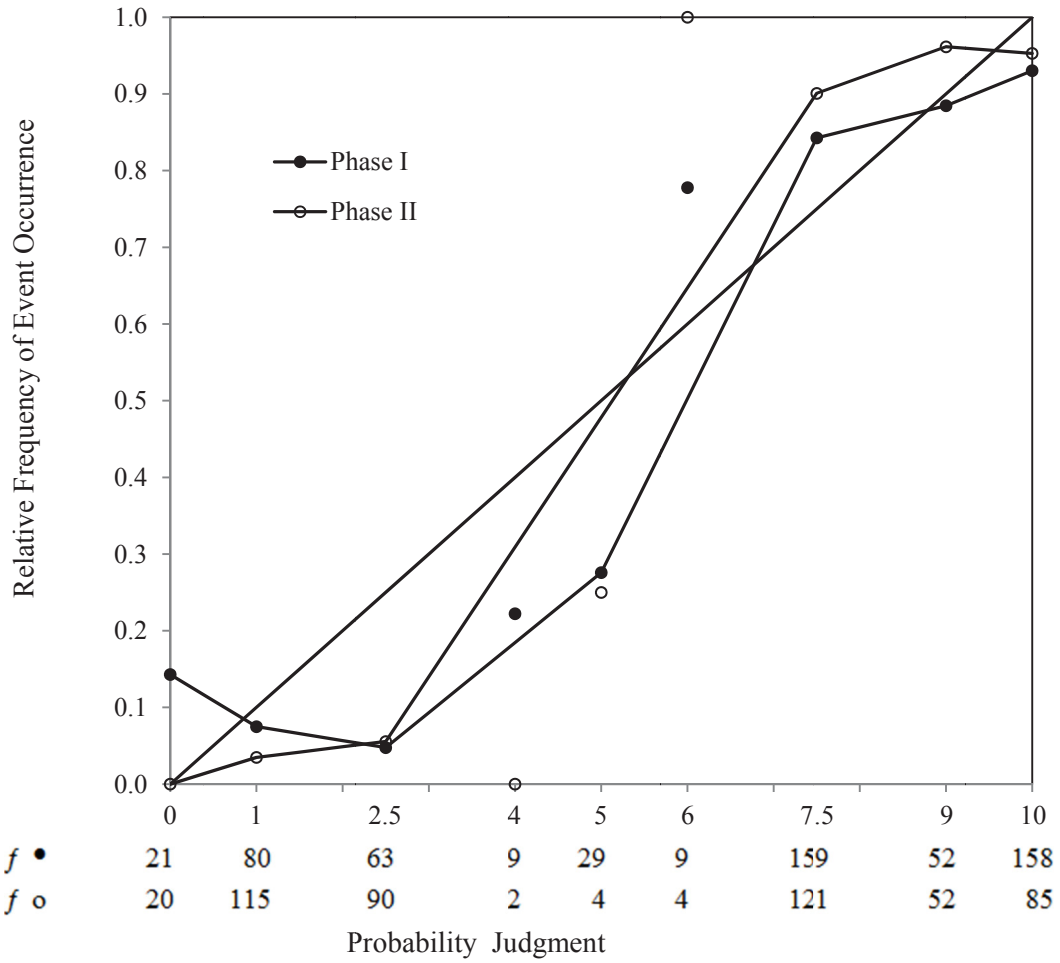

Figure 4: Reliability Diagram for the MEA Division by Phase

4.1.2 Analyst Experience

Junior and senior analysts showed comparable degrees of calibration. However, senior analysts had much better discrimination, accounting for 76% of total outcome variance in comparison to about 59% for junior analysts. In percentage-correct terms, this difference was reflected in about a five percentage-point advantage for senior analysts on both liberal and conservative measures. Figure 5 shows that senior analysts were more accurate at the extremes, especially in forecasting definitive non-occurrences (bearing in mind the small sample size within the 0/10 judgment category).

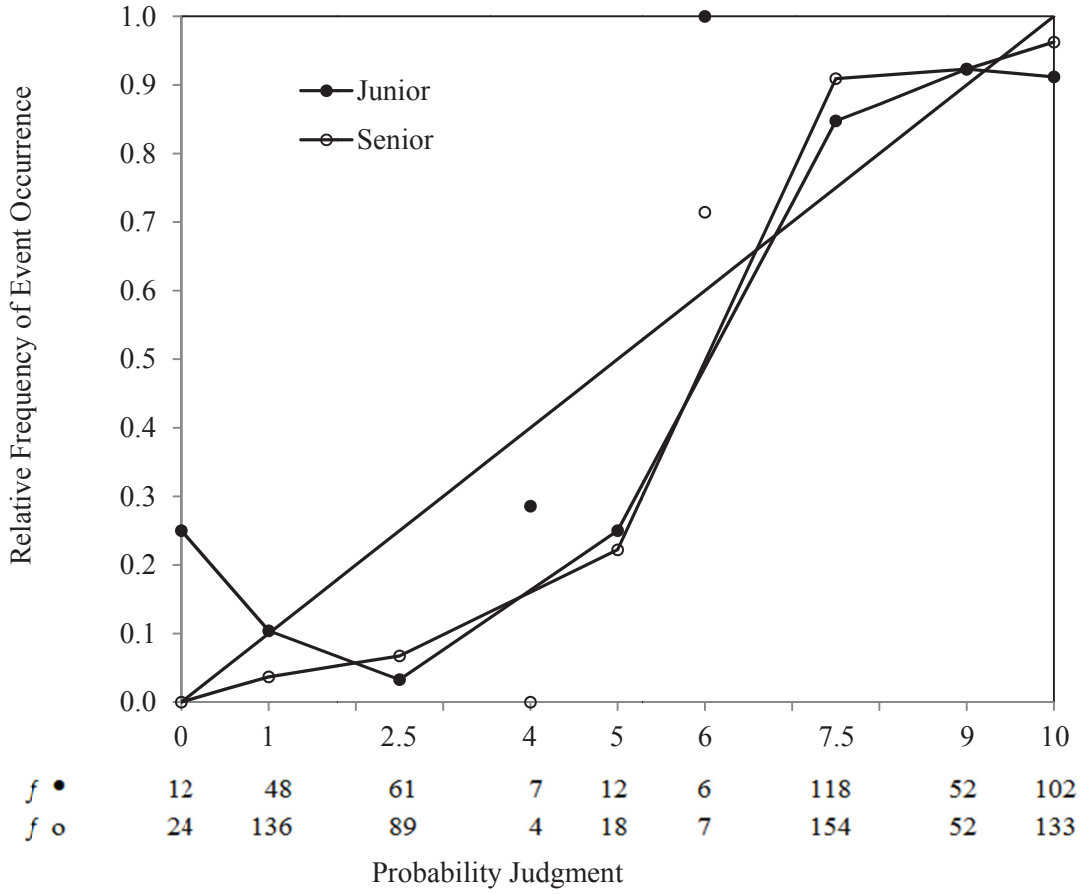

Figure 5: Reliability Diagram for the MEA Division by Analyst Experience

There was no evidence to support claims that experience tends to inflate confidence without improving accuracy. Indeed, the calibration curves for senior and junior analysts track one another very closely, with both reflecting the aforementioned pattern of underextremity bias, characteristic of overly conservative forecasting.

Finally, it is noteworthy that the small sample of multiple-analyst predictions showed substantially inferior performance. First, the relatively greater discrepancy between liberal and conservative percentage-correct measures in the multiple-analyst group implies that team judgments produced a greater proportion of fence-sitting verdicts, which were punishable under the conservative criterion. Based on that criterion, that translated into a 10-point spread in favour of junior analysts who predicted alone and about a 15-point spread in favour of senior analysts who predicted alone.

4.1.3 Forecast Difficulty

Although calibration was identical across difficulty levels, easier forecasts showed superior discrimination. Whereas 81% of total outcome variance was accounted for by analysts' forecasts in the easier subset, that figure dropped to 61% in the harder subset. The spread was also evident in the measures of percentage correct. The calibration curves by judgment difficulty are shown in Figure 6.

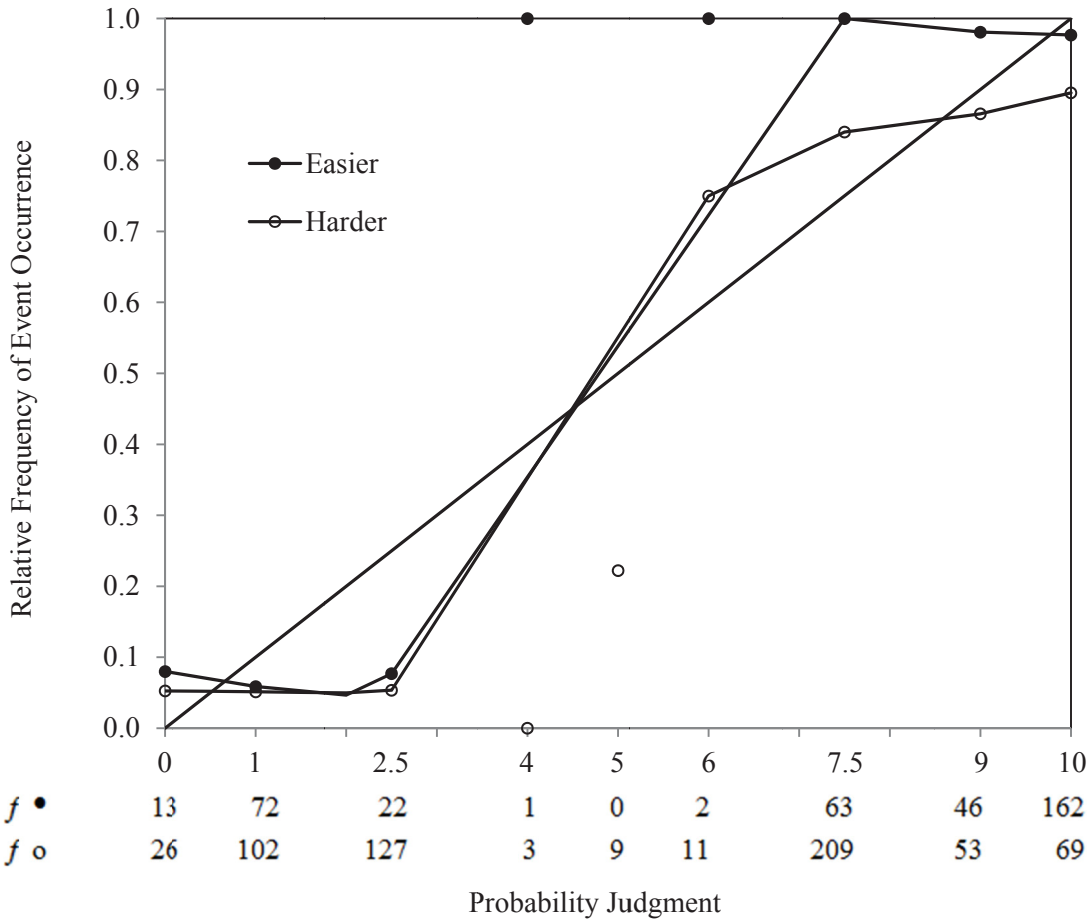

Figure 6: Reliability Diagram for the MEA Division by Forecast Difficulty

4.1.4 Forecast Importance

About three-quarters of the overall set of MEA Division forecasts were coded as being of moderate to high importance. As Table 5 shows, those forecasts deemed to be of higher importance were better in terms of both calibration and discrimination as compared with those deemed to be of low to moderate importance. Figure 7 shows the calibration curves by forecast importance.

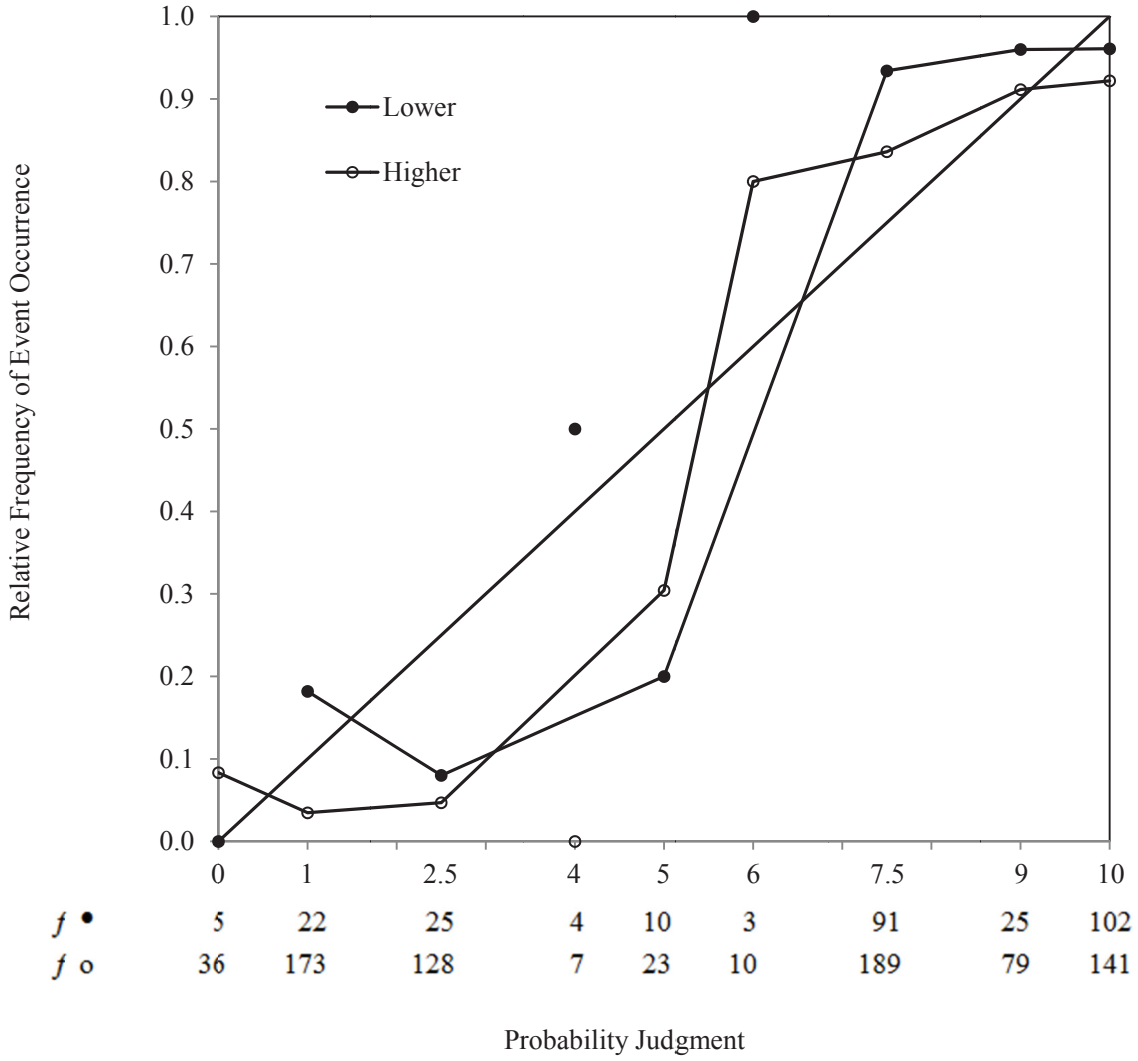

Figure 7: Reliability Diagram for the MEA Division by Forecast Importance

4.1.5 Forecast Prominence

As both Table 5 and Figure 8 reveal, key forecasts identified in the equivalent of the executive summaries of the reports showed slightly better judgment quality in terms of both calibration and discrimination as compared to forecasts reported in the main body of the report.

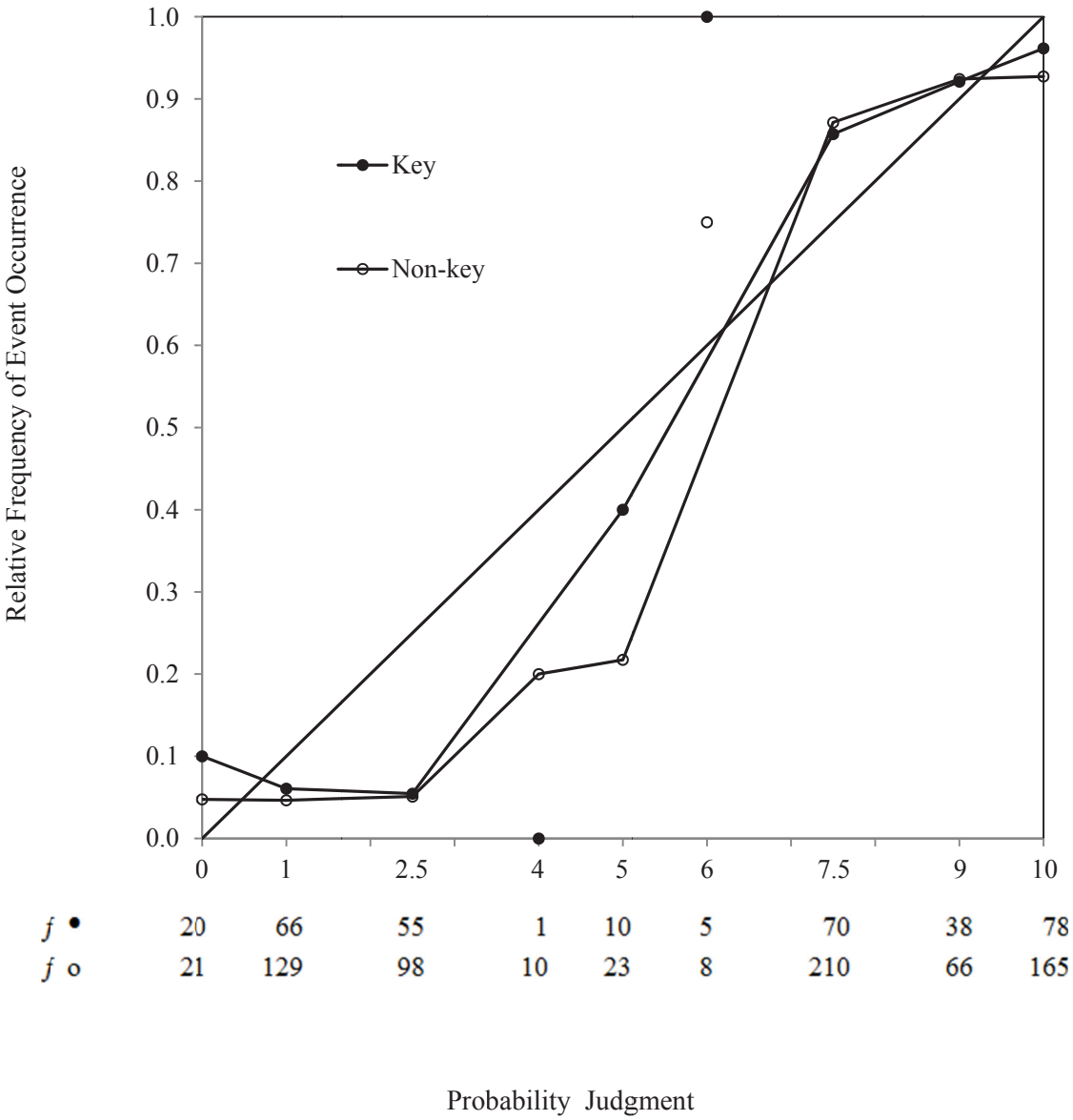

Figure 8: Reliability Diagram for the MEA Division by Forecast Prominence

4.1.6 Report Origin

There was a slight trade-off between calibration and discrimination in terms of report origin. As Table 5 shows, forecasts from externally tasked reports had about a five-point advantage over internally tasked reports in terms of explaining total outcome variance (about 82% vs. 77%, respectively). However, calibration was slightly better for forecasts in internally tasked reports. Overall, the difficulty-adjusted "skill" score was higher for forecasts in externally tasked reports.

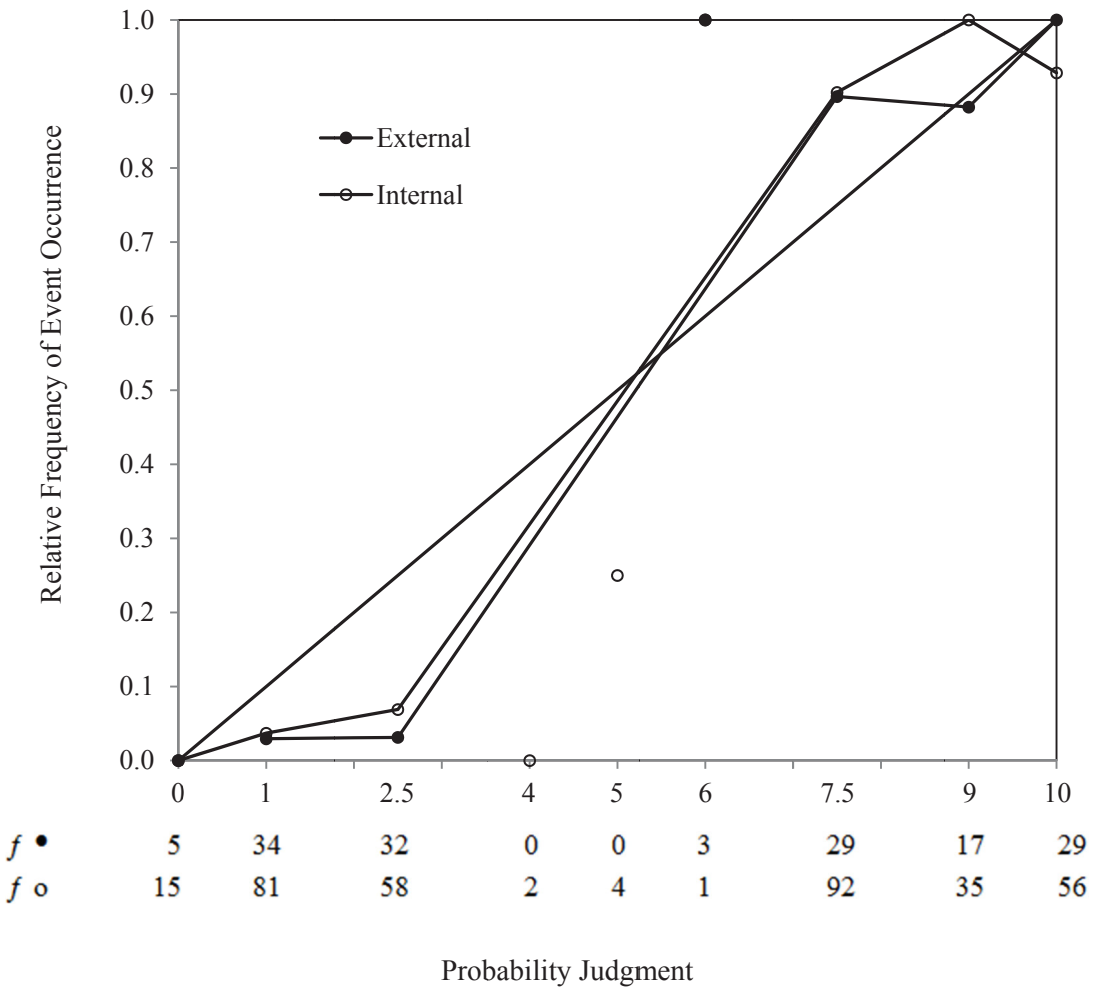

Figure 9: Reliability Diagram for the MEA Division by Report Origin

4.1.7 Region of Interest

Forecast quality for the Middle East and Africa were comparable. Calibration was virtually identical across regions, whereas discrimination was a little better for forecasts about Africa.

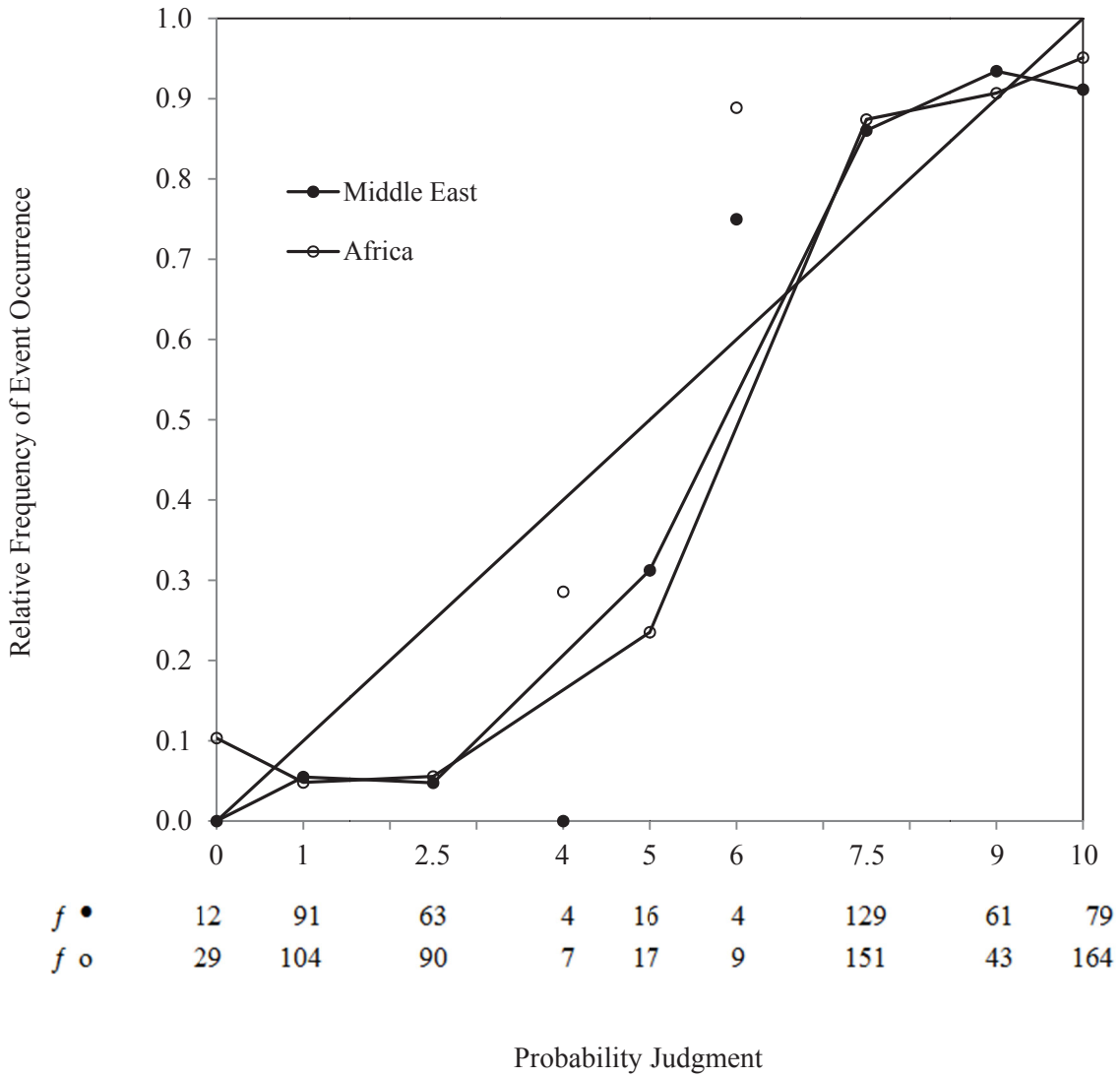

Figure 10: Reliability Diagram for the MEA Division by Region of Interest

4.1.8 Forecast Time Frame

As Table 5 shows, the vast majority of forecasts were made over a three-month to one-year time frame. Calibration was best for forecasts made over time frames greater than one month and less than six months. Discrimination was best for forecasts made over a time frame greater than one year, where in fact all 17 forecasts were correctly classified. More generally, discrimination increased monotonically from the shortest time frame of less than one month to the longest time frame of more than one year. Forecasting skill also increased over time frame, although not in a perfectly monotonic function.

4.2 Inferred Probabilities From Outside the MEA Division

To begin, some comparisons with the results of the preceding section may be drawn. First, note that the sample sizes between the two sets of analyses differ by roughly a factor of three. The inferred probability sample is quite small for a calibration study and must be interpreted with caution, not only because the numeric probabilities were inferred by a single coder but also because of the small sample size.

Moreover, the two forecast samples differ in terms of the base rate of event occurrence by just over a 10-point margin. Whereas about 57% of the forecasted events occurred in the MEA Division sample, about 68% did so in the inferred probability sample. As the base rate approaches the extremes, total outcome variance decreases, and the task of forecasting becomes easier (Tetlock, 2005). It is important to bear this in mind when drawing comparisons between the samples.

The adjusted measure of discrimination η^{2*}, however, controls for this by representing the proportion of the total variance that is explained by assignments to the judgment categories. As well, the difficulty-adjusted probability score PS^* takes this into account.

Bearing these caveats in mind, the findings indicate a high level of forecasting quality in the inferred probability sample. Overall, as Table 6 shows, 95% of forecasts were correctly classified. Moreover, the identical value for liberal and conservative measures of percentage correct means there were no inferred fence-sitting forecasts of fifty-fifty.

Calibration was the same as the level reported for the MEA Division sample. Discrimination (as measured by η^{2*}) was excellent and exceeded the MEA Division level by about 11 percentage points (roughly 79% in this sample vs. 68% in the MEA Division sample).

Forecaster skill was somewhat lower for Phase II forecasts from the MEA Division ($PS^* = .61$) than from elsewhere ($PS^* = .70$). This comparison is informative because all Phase II forecasts were scored by the same coder and relied on the same procedure for handling indeterminate probabilities. However, as was shown in Table 1, an additional 10% of the forecasts made in non-MEA Division Phase II reports were indeterminate probabilities (i.e., 22% in non-MEA Division reports vs. 12% in MEA Division reports). Such differences in forecast quality are not captured by the metrics reported here but must nevertheless be borne in mind when interpreting the results.

Table 6: Summary Statistics for Inferred Probabilities from Outside the MEA Division

	N	$\bar{d}$	%CP_L	%CP_C	*PS*	VI	CI	DI	η^{2*}	PS^*
Overall	349	.68	95.1	95.1	.059	.219	.014	.174	.791	.700
Analyst experience:										
Junior	35	.71	91.4	91.4	.084	.204	.014	.135	.558	.525
Senior	180	.72	95.0	95.0	.055	.201	.012	.158	.776	.684
Multiple	134	.60	96.3	96.3	.057	.239	.023	.206	.852	.747
Difficulty:										
Easier	170	.75	98.8	98.8	.026	.186	.015	.175	.936	.769
Harder	179	.60	91.6	91.6	.089	.239	.018	.168	.688	.617
Importance:										
Lower	37	.86	97.3	97.3	.039	.117	.015	.093	.734	.505
Higher	312	.65	94.9	94.9	.061	.226	.016	.181	.795	.710
Prominence:										
Key	76	.63	97.4	97.4	.039	.233	.016	.210	.890	.826
Non-key	273	.69	94.5	94.5	.064	.214	.015	.165	.763	.658
Origin:										
External	110	.71	96.4	96.4	.049	.206	.016	.173	.828	.723
Internal	239	.66	94.6	94.6	.063	.224	.016	.177	.783	.689
Source:										
IAS	215	.72	94.4	94.4	.060	.201	.011	.152	.748	.658
Interdepartmental	134	.60	96.3	96.3	.057	.239	.023	.206	.852	.747
Time frame:										
Up to 1 month	2	.50	100.0	100.0	.063	.250	.063	.250	1.00	.750
2 - 3 months	5	.60	80.0	80.0	.125	.240	.125	.240	1.00	.512
3 - 6 months	59	.68	96.6	96.6	.055	.218	.027	.191	.854	.718
6 months - 1 year	200	.69	94.5	94.5	.064	.216	.014	.165	.755	.661
> 1 year	81	.65	96.3	96.3	.043	.226	.020	.203	.887	.793

Figure 11 reveals that the characteristic pattern of underextremity bias evident in the MEA Division sample is also observed in the inferred probability sample. Indeed, the calibration curve indicates a step function where low-probability forecasts have about a 5% rate of outcome occurrence and high-probability forecasts have about a 95% rate.

In the following subsections, the findings from the inferred probability sample are analyzed and discussed in terms of a subset of the variables reported in Table 6. Reliability diagrams are presented for judgment difficulty, report origin, and forecast source only because the other variables have one level with too few cases to report a meaningful visual representation of the variables' effects. The fuller set of variables is explored in the combined analysis reported in Section 4.3.

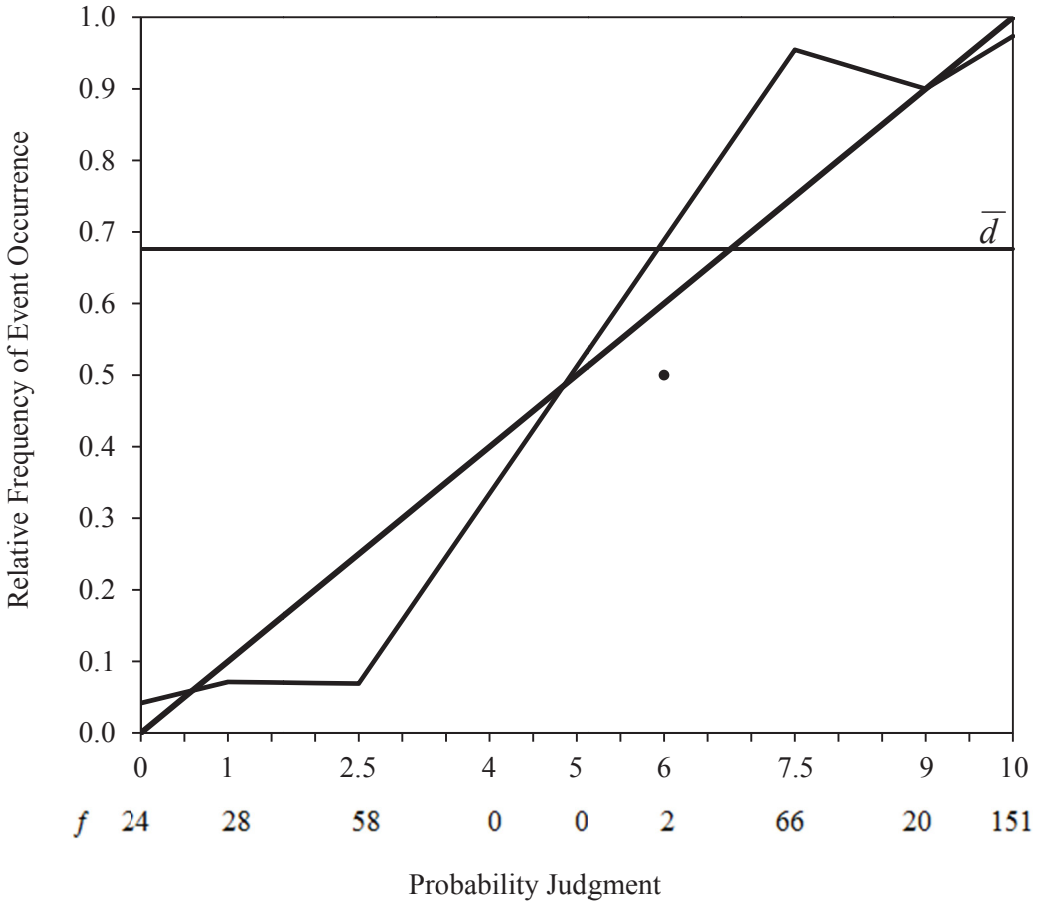

Figure 11: Reliability Diagram for the Overall Inferred Probability Sample

4.2.1 Forecast Difficulty

Forecasts coded as relatively easier to make presented exceptionally well. Over 99% were correctly classified, and about 94% of total outcome variance was accounted for by analysts' forecasts (versus about 69% for the harder forecasts). Calibration was also very good and somewhat better than for the harder forecasts. As Figure 12 shows, the easier forecasts conformed to an almost perfectly discriminating step function.

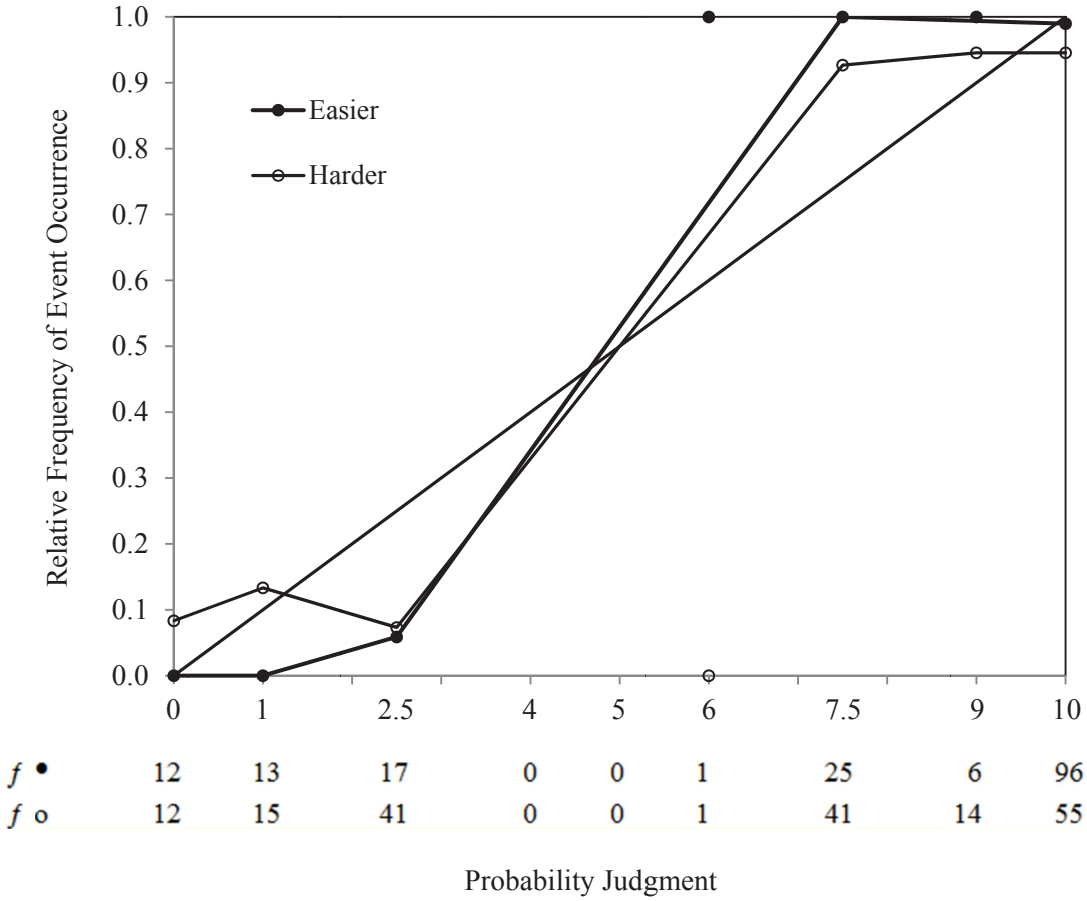

Figure 12: Reliability Diagram for Forecast Difficulty in the Inferred Probability Sample

4.2.2 Report Origin

As with the MEA Division sample, forecasts from externally tasked reports had about a five-point advantage over internally tasked reports in terms of explaining total outcome variance (about 83% vs. 78%, respectively). Calibration did not differ as a function of report origin.

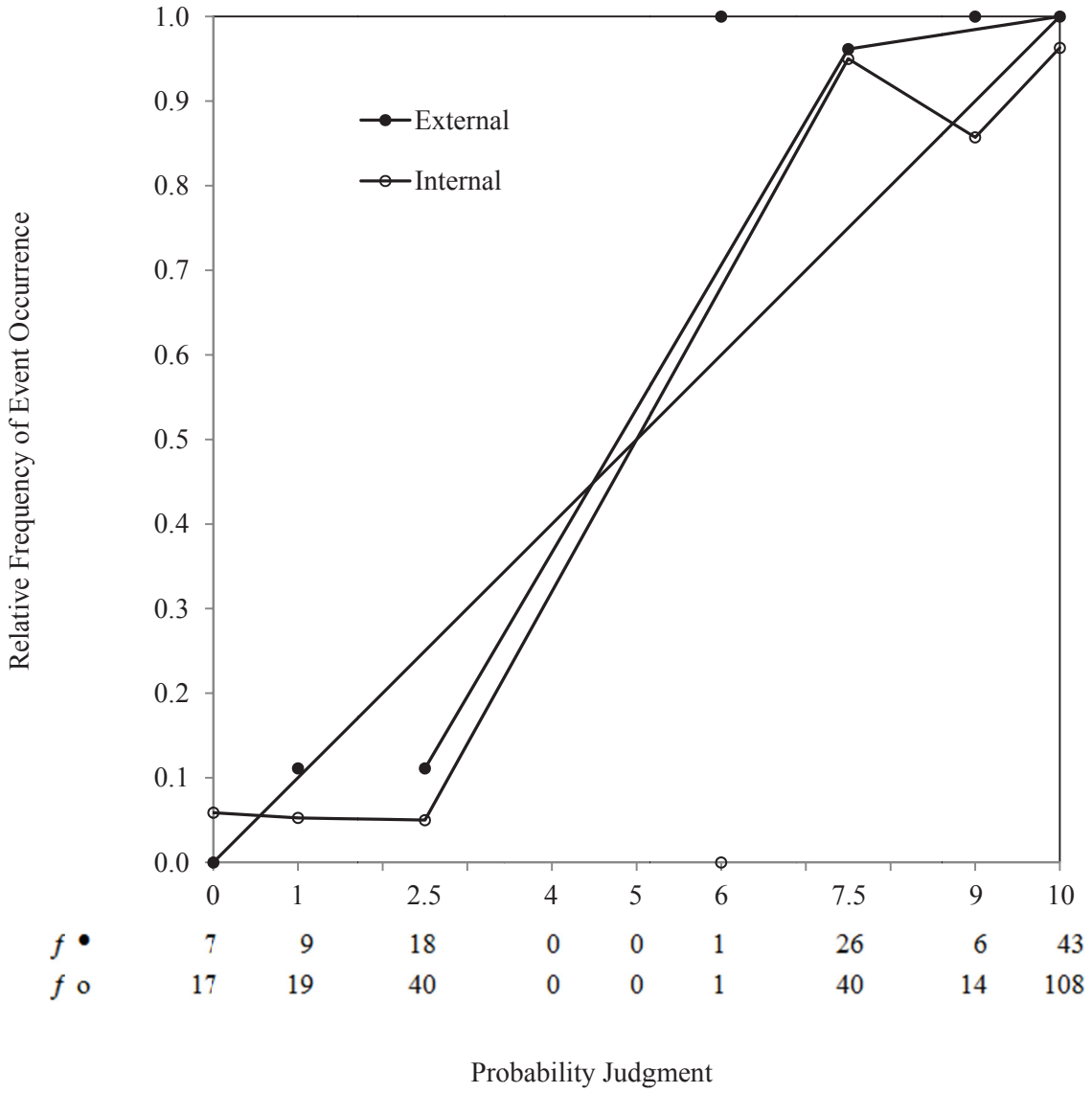

Figure 13: Reliability Diagram for Report Origin in the Inferred Probability Sample

4.2.3 Forecast Source

IAS forecasts showed substantially better calibration than interdepartmental forecasts. However, interdepartmental forecasts showed a roughly 10-point advantage over IAS forecasts in terms of proportion of outcome variance explained (about 85% vs. 75%, respectively). Probability scores and percentage correct measures were comparable across source. However, difficulty-adjusted probability scores showed an advantage for interdepartmental forecasts. This difference reflects the fact that the VI is closer to maximal difficulty in the interdepartmental set of forecasts (VI = .24) than in the IAS set (VI = .20). Figure 14 shows the calibration curves of the inferred probability sample as a function of forecast source.

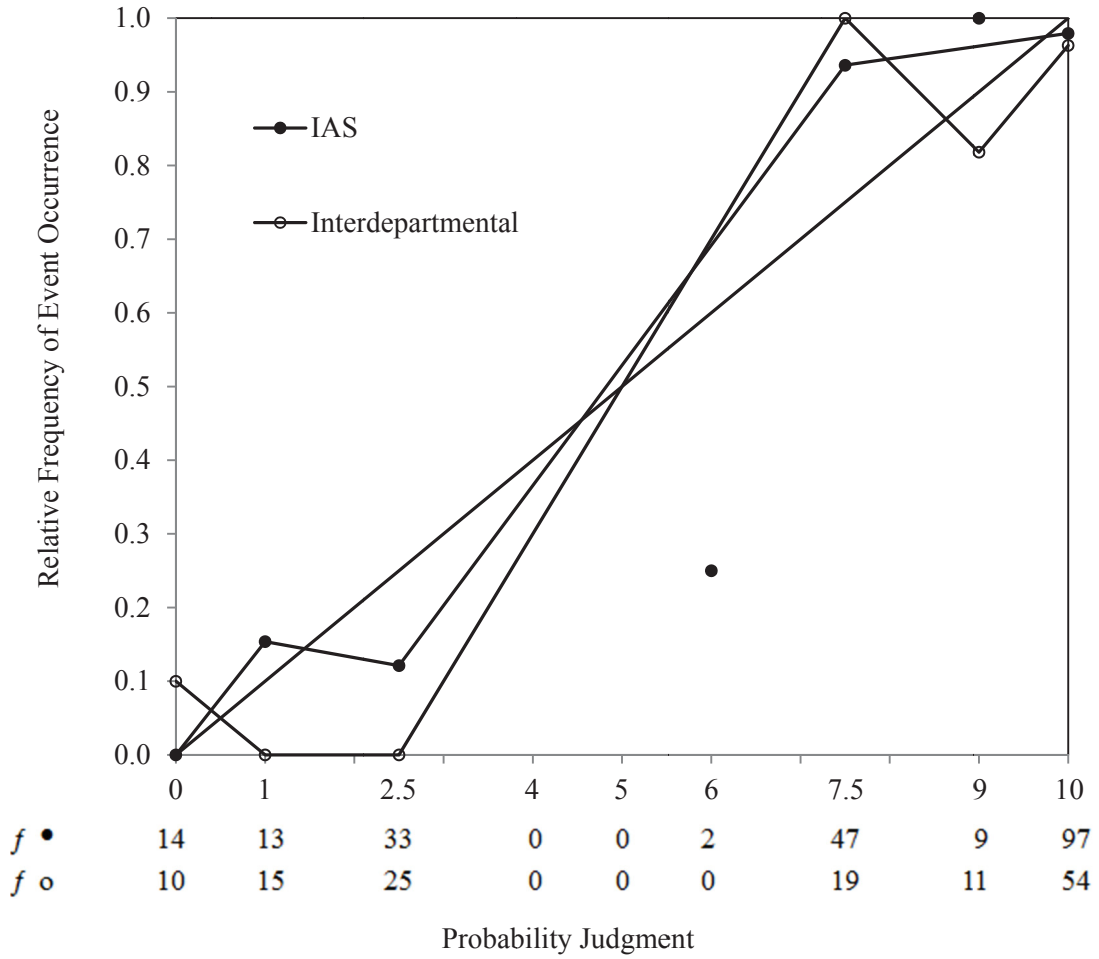

Figure 14: Reliability Diagram for Forecast Source in the Inferred Probability Sample

4.3 Combined Analysis

Combining forecasts from the samples using numeric and inferred probability qualifiers was by no means a foregone conclusion at the start of Phase II. As noted earlier, the two samples differ in important ways that ought not to be obscured even by commonalties in results. In the MEA Division, analysts generated numbers for their forecasts. In all other cases, Coder II inferred the numbers. Thus, the results obtained from the inferred sample must be interpreted with added caution. Nevertheless, there are some striking similarities between the two samples, and these suggest the value of presenting a combined analysis if for no other reason than to benefit from a larger sample with more reliable estimates of the performance measures.

As Table 7 shows, the base rate of event occurrence within the overall sample was 59%. Over 90% of outcomes were correctly classified by both liberal and conservative measures. Unsurprisingly, the pattern of underextremity bias with some local overconfidence at the extremes was evident (see Figure 15).

Table 7: Summary Statistics for the Combined Sample

	N	$\bar{d}$	$\%CP_L$	$\%CP_C$	*PS*	VI	CI	DI	η^{2*}	PS^*
Overall	1422	.59	92.7	90.5	.083	.241	.013	.171	.708	.641
Analyst Experience:										
Junior	453	.63	89.6	87.2	.110	.232	.016	.138	.589	.499
Senior	797	.58	94.2	92.1	.070	.243	.014	.187	.769	.703
Multiple	172	.54	93.5	91.9	.074	.248	.014	.189	.750	.698
Difficulty:										
Easier	551	.74	96.7	96.7	.044	.195	.015	.165	.847	.693
Harder	788	.52	89.5	88.5	.108	.250	.015	.157	.626	.560
Importance:										
Lower	324	.78	93.6	90.7	.078	.173	.019	.114	.650	.447
Higher	1098	.54	92.4	90.4	.085	.248	.013	.177	.709	.655
Prominence:										
Key	419	.56	93.4	91.2	.077	.246	.013	.182	.734	.683
Non-key	1003	.61	92.3	90.2	.086	.238	.014	.167	.696	.621
Origin*:										
External	259	.59	94.3	93.7	.056	.242	.017	.202	.831	.759
Internal	583	.58	95.8	95.8	.069	.244	.015	.190	.777	.708
Time frame*:										
Up to 1 month	21	.57	90.5	90.5	.103	.245	.035	.177	.555	.549
2 - 3 months	38	.55	92.1	92.1	.074	.247	.010	.182	.668	.702
3 - 6 months	237	.51	94.1	94.1	.070	.250	.015	.195	.772	.719
6 mo. - 1 year	446	.62	95.0	94.2	.065	.236	.017	.188	.792	.706
> 1 year	98	.59	96.9	96.9	.040	.242	.021	.222	.913	.830

*Phase II only.

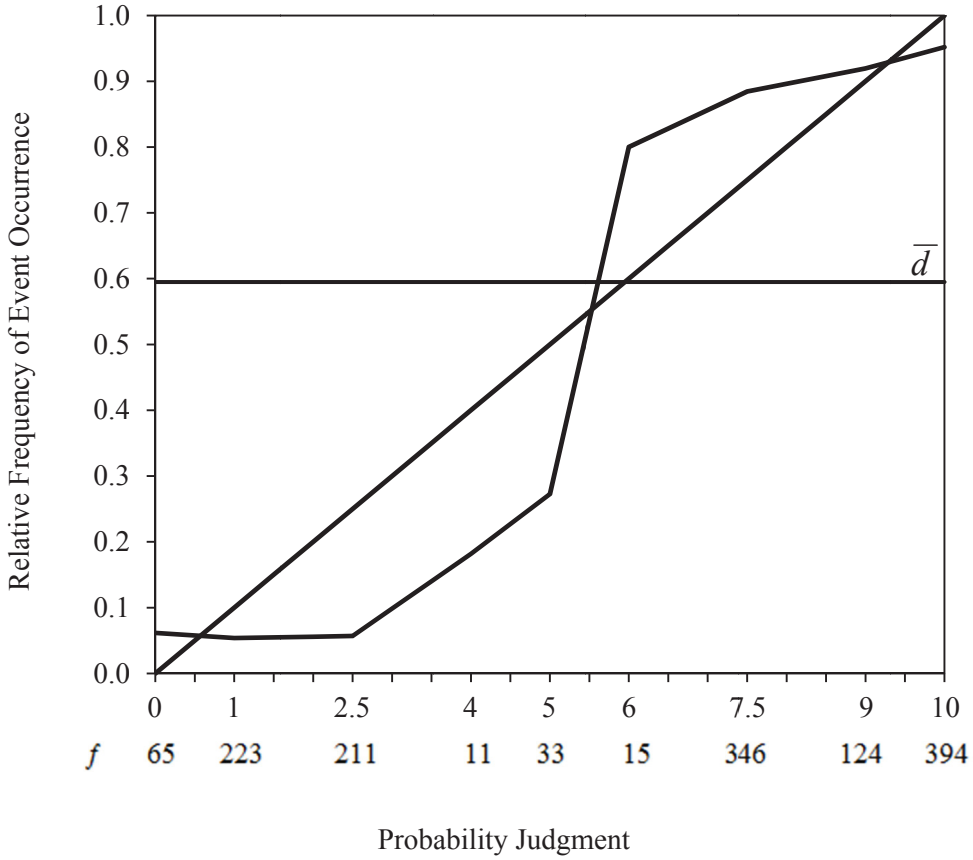

Figure 15: Reliability Diagram for the Overall Combined Sample

4.3.1 Analyst Experience

Analyst experience showed little effect on calibration. However, discrimination was far better (by about eight percentage points) among the senior analysts. As Figure 16 shows, senior analysts were better calibrated than junior analysts at the extremes.

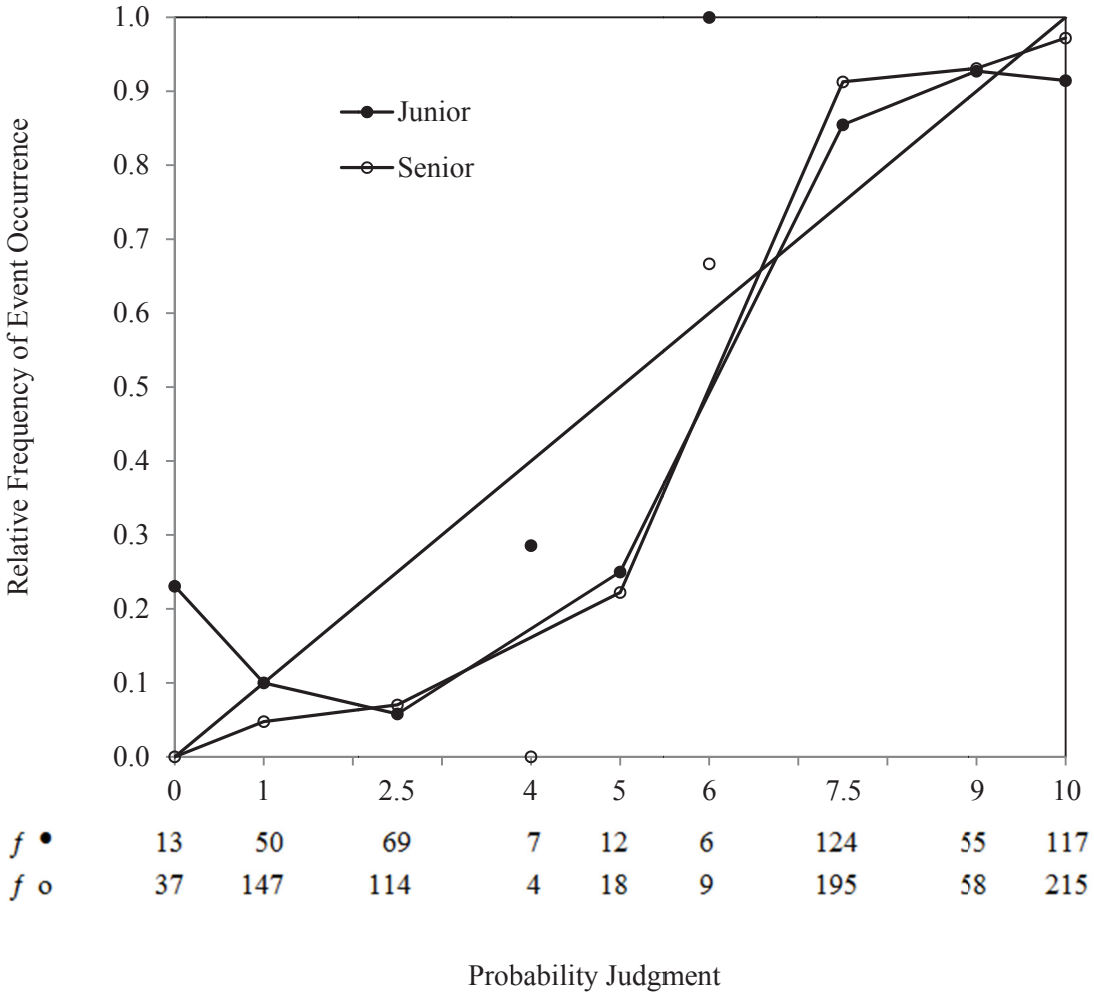

Figure 16: Reliability Diagram for Analyst Experience in the Combined Sample

4.3.2 Forecast Difficulty

Although calibration was indistinguishable between easier and harder forecasts, forecast difficulty had a large effect on discrimination. Whereas easier forecasts explained about 85% of outcome variance, harder forecasts explained only about 63%. It is evident from Table 7 that the harder forecasts were in fact more variable: the base rate was about 74% for the easier forecasts and about 52% for the harder forecasts, the latter being close to maximum uncertainty. This is reflected in the difficulty-adjusted probability scores, which show a distinct skill advantage for the senior analysts. Moreover, Figure 17 shows that the discrimination success for easier forecasts was driven in large measure by the large number of correctly forecasted 10/10 cases.

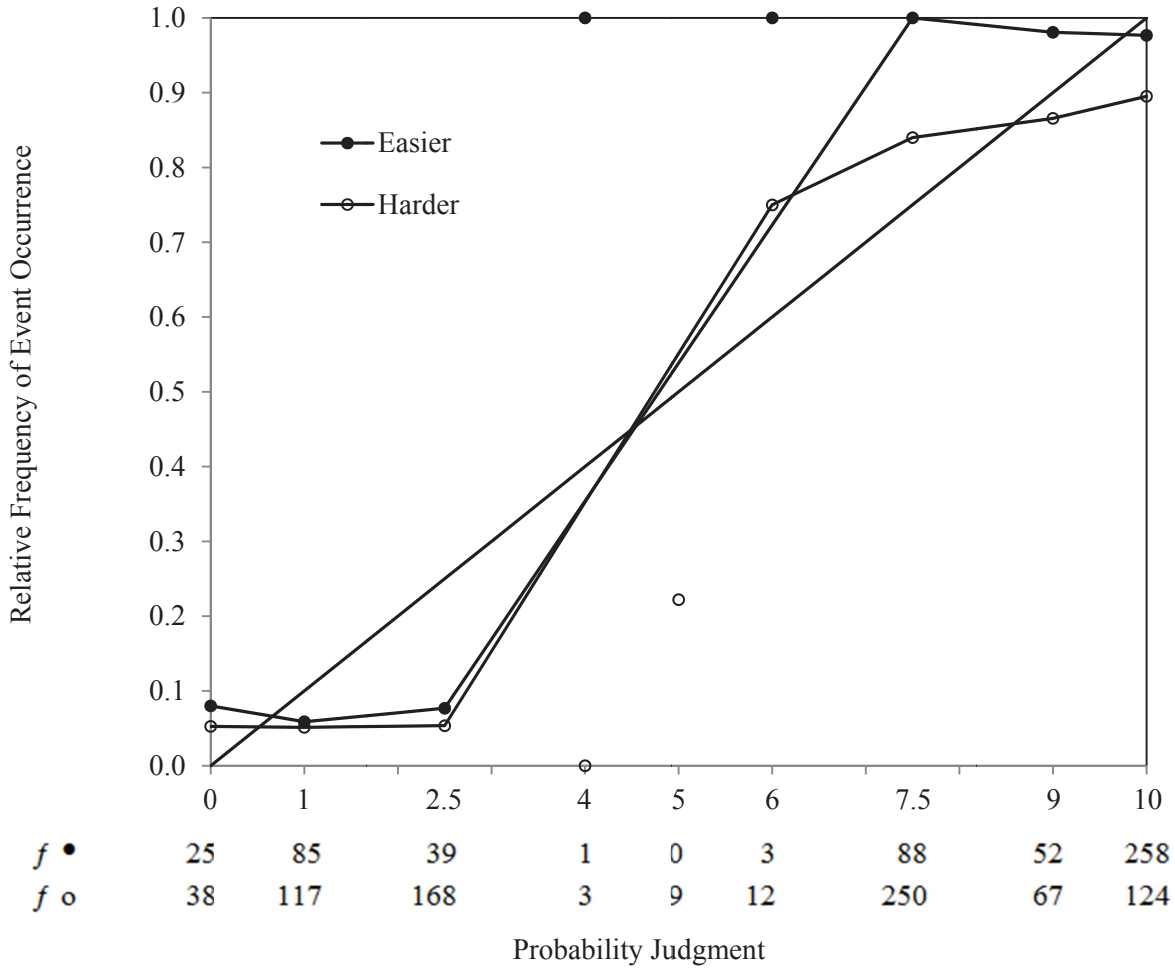

Figure 17: Reliability Diagram for Forecast Difficulty in the Combined Sample

4.3.3 Forecast Importance

Forecasts deemed to be of higher importance were better than forecasts deemed to be of lower importance in terms of both calibration and discrimination. This advantage was evident in spite of the fact that outcome variance was greater in the higher-importance subset. The calibration curves are shown in Figure 18.

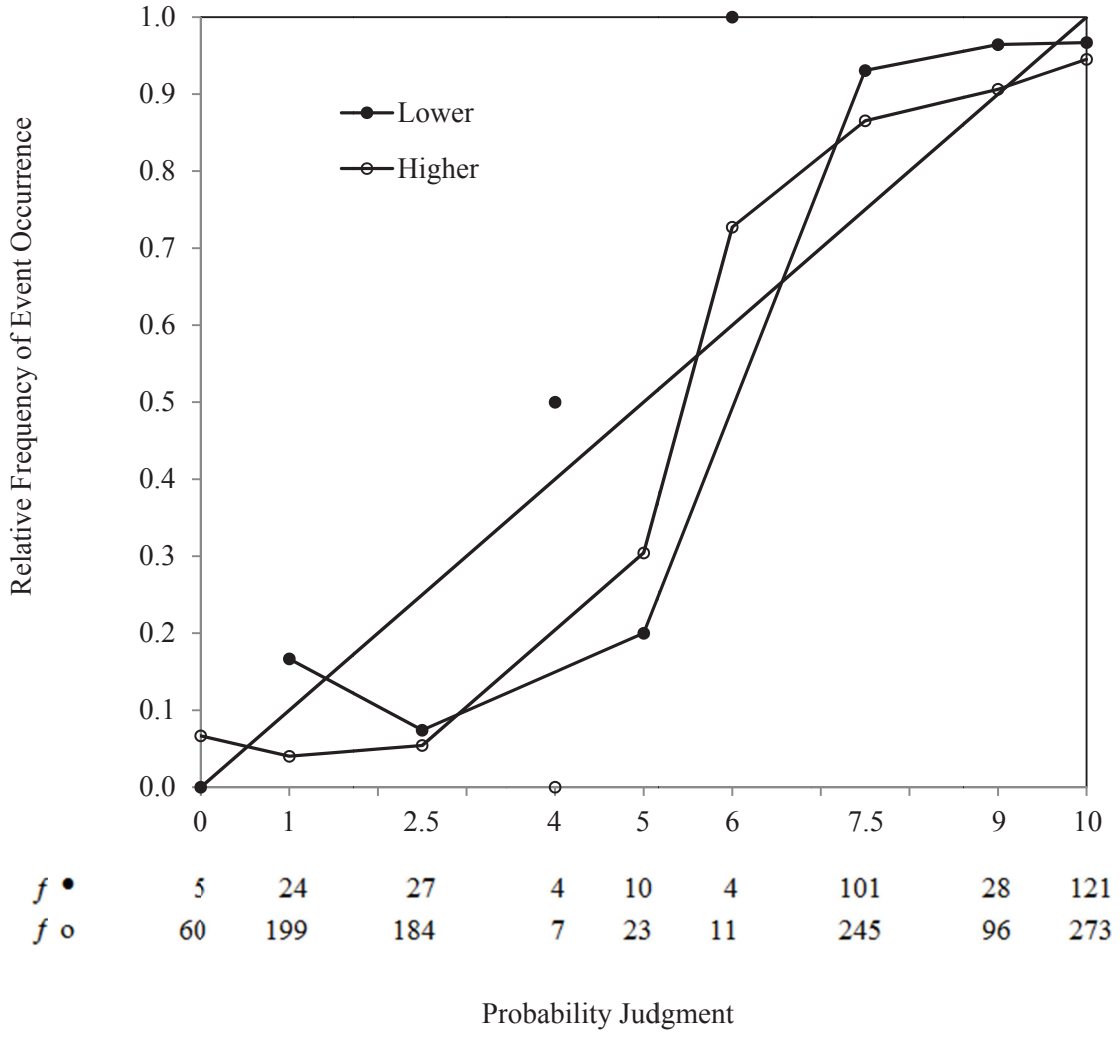

Figure 18: Reliability Diagram for Forecast Importance in the Combined Sample

4.3.4 Forecast Prominence

Forecasts that appeared in key judgments had slightly better calibration and discrimination than forecasts that appeared in the main body of intelligence reports. The calibration curves for the two sets of forecast are shown in Figure 19.

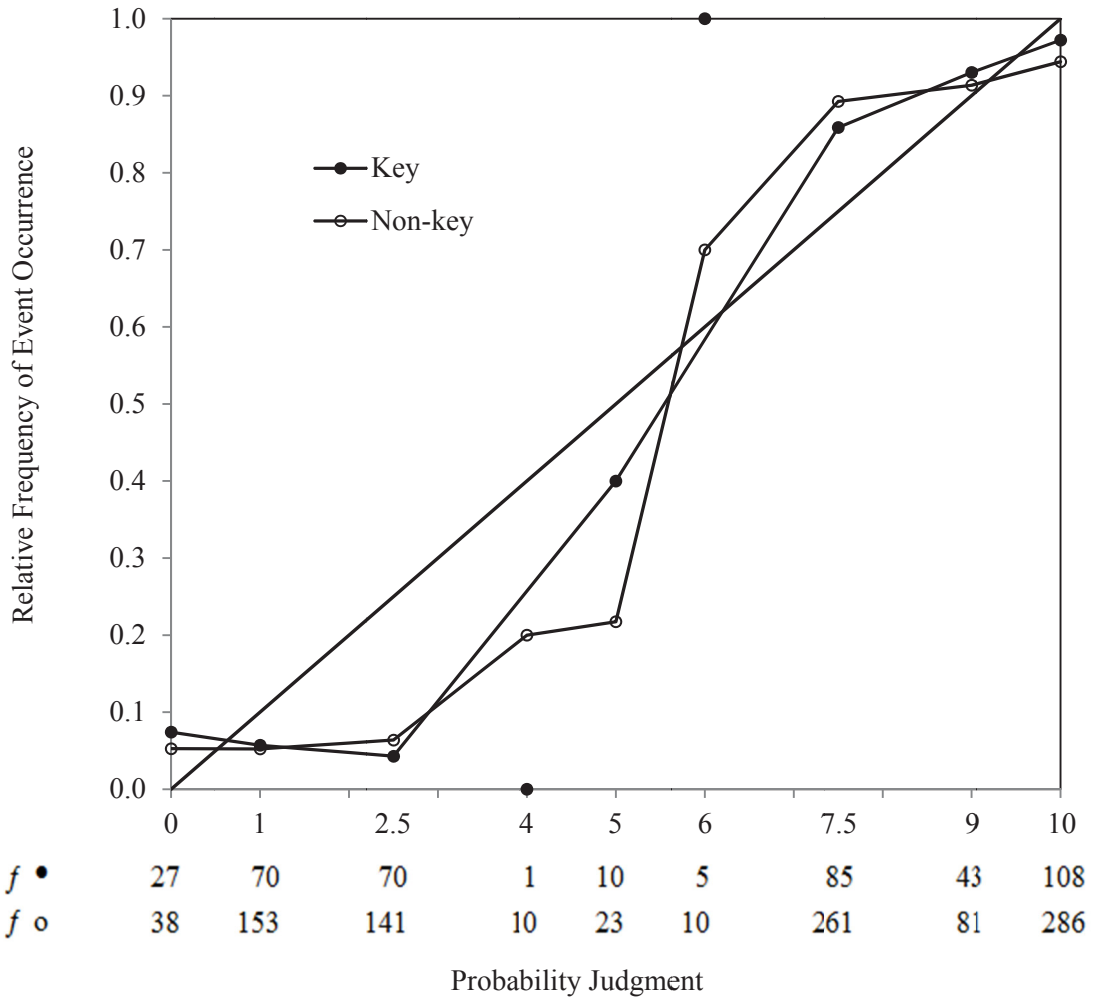

Figure 19: Reliability Diagram for Forecast Prominence in the Combined Sample

4.3.5 Report Origin

Overall, externally tasked reports included forecasts that were slightly less well calibrated than forecasts from internally tasked reports but which had somewhat better discrimination. The benefit of better discrimination arguably outweighed the slightly poorer calibration given that the externally tasked reports yielded a more impressive skill score (PS^* = .76 for external and .71 for internal).

Figure 20 reveals that forecasts in externally tasked reports were perfectly accurate at both extremes. Thus, the local overconfidence exhibited in the overall sample was not attributable to forecasts made in externally tasked reports. Indeed, it is a striking finding that all 84 forecasts made with certainty (i.e., where the forecasts were either 0/10 or 10/10) in externally tasked reports were correct.

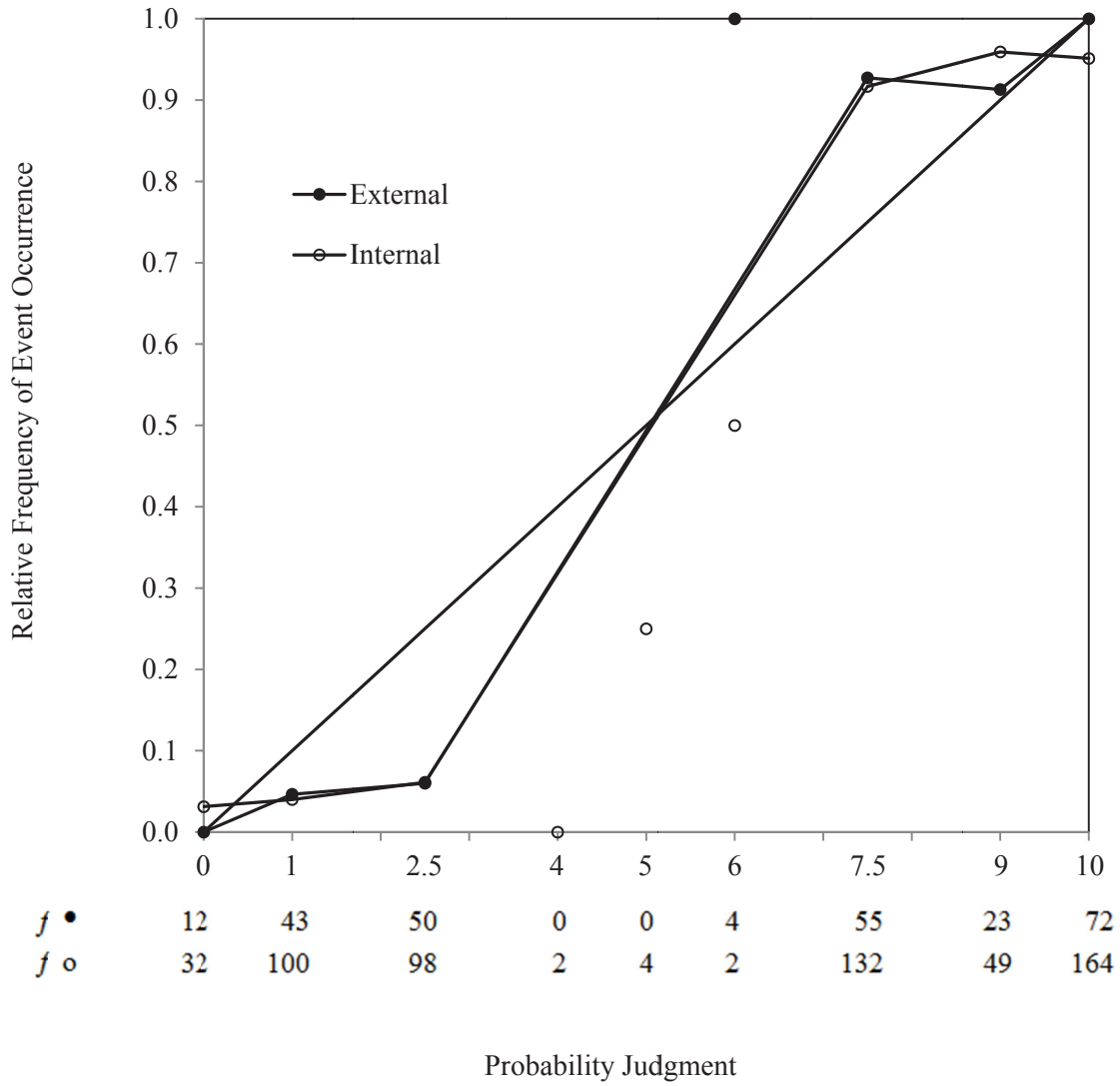

Figure 20: Reliability Diagram for Report Origin in the Combined Sample

4.3.6 Forecast Time Frame

As shown in Table 7, forecasting quality was sensitive to the forecast's time frame. As in the MEA Division analysis, discrimination increased monotonically with time frame. In contrast, but for the shortest time frame of less than one month (where there were relatively few cases), calibration decreased monotonically with time frame. Thus, there appears to be a trade-off between calibration and discrimination that varies by time frame. Overall, however, forecasting skill increased with time frame, with PS^{*} = .55 for the shortest time frame and PS^{*} = .83 for the longest time frame. Thus, forecasting quality in this sample of forecasts appears to have benefitted from longer time frames over which the forecasted events could occur.

5 Discussion

The present field study examined the quality of probabilistic forecasts made in Canadian strategic intelligence assessments. Quality was measured using quantitative scoring rules commonly employed in other areas of expert forecast monitoring and also in basic and applied research on the accuracy of probability and confidence judgments.

The study revealed many important and novel findings of relevance both to the applied behavioural science literature on judgment accuracy and to the intelligence community. Perhaps the most obvious starting place for summarizing those findings is in terms of the overall level of performance witnessed, which was surprisingly high and robust across factors that one might otherwise have thought would produce greater variability in the findings. For instance, a high level of forecasting accuracy, both in terms of discrimination and calibration, was evident across the MEA Division sample that produced numeric forecasts and the remaining sample for which numeric values had to be inferred from verbal expressions of uncertainty. Forecasting accuracy was high not only for IAS forecasts but also for those produced by interdepartmental committees.

Indeed, high forecasting quality was evident in every factor-level analysis of the overall sample of 1,422 forecasts shown in Table 7, as evidenced by difficulty-adjusted skill scores ranging from .45 to .83. (Recall that a skill score of 0 represents no improvement over merely forecasting the relevant base rate of event occurrence [i.e., assigning a probability equal to $\bar{d}$ in the relevant row of Table 7], while a skill score of 1 reflects perfect "crystal ball" forecasting skill.) Perhaps in more intuitive, if less refined, terms, one observes as well from Table 7 that the percentage of correct classifications is above 90% even by the agnostic-punishing, conservative measure for all factor-level analyses except that for junior analysts, where the value dipped to a nonetheless highly respectable 87%.

Forecast quality could also be assessed against a variety of mindless strategies. For instance, Tetlock (2005) compares the forecasting performance of political science experts to a simple, random guessing strategy where one guesses the long-range average based on the assumption of equiprobable alternatives. Tetlock describes this as a "chimp" strategy since it involves mindless guessing. In the present study, that long-range average is .50 since there are only two possible outcomes. Tetlock's chimp strategy (Chimp I) yields a poor *PS* score of .25 as compared with *PS* = .08 for the overall sample. We tried two other chimp strategies. In Chimp II, we randomly picked probabilities of either 1 (10/10) or 0 (0/10). This yielded an even worse *PS* of .53. Clearly, chance was not on our side in this random draw. In Chimp III, we randomly picked one of the nine probability levels used in the study. This yielded a *PS* score of .35, falling between the scores of Chimp I and Chimp II. In short, analysts' forecast quality was not even in the same ballpark.

However, this comparison to the lowest common denominator should offer little comfort. One would certainly hope—and, indeed, expect—analysts to do better at forecasting than what would be achieved by random guesswork. However, even if we compare performance in the present study to the performance of expert human forecasters and case-specific models in Tetlock's (2005) study, the results are impressive. Among Tetlock's expert sample, CI was .025 and DI was .024. Tetlock further reported that "the best human forecasters were hard-pressed to predict more than 20 percent of the total variability in outcomes (using the D/VI "omniscience" index)" (p. 53). In Tetlock's study, the human experts were out-forecasted by both case-specific extrapolation algorithms and generalized autoregressive distributed lag models, which respectively accounted for 25 to 30 percent and 47 percent of the outcome variance. In the present study, by comparison,

analytic forecasts explained 71 percent of the outcome variance, far exceeding the performance of both expert human and model forecasts witnessed in Tetlock's study.

This naturally raises the question of why analysts' forecasts appear to be so good. One possibility is that they are in fact as good as they appear, in which case we might try to better account for why they are so good, even while feeling reassured that strategic intelligence forecasting quality is high (at least in terms of the scoring rules employed in this study). Another, less desirable, possibility, however, is that forecasting quality in this study was artificially inflated by certain methodological or data management decisions. We cannot simply rule out that possibility, given that there are some plausible hypotheses to consider. First, in terms of isolating the set of forecast cases, 213 out of 1,934 forecasts were indeterminate (i.e., assigned an "X/10") and were excluded from further consideration (leaving only 1,721 determinate forecasts). In Phase I, however, indeterminate forecasts were assigned a probability of 40% to 60%. Thus, one could examine the effect of reintroducing those forecasts and substituting a probability of .5 (i.e., 50%). Likewise, the statistical analyses conducted excluded forecasts for which there was not a clear "occurred" or "did not occur" outcome. In some cases, coders assigned a partial "occurred" or a partial "did not occur" code to the relevant outcome, while, in others, the wording of the judgment was such that it could not be falsified, or the coder was unable to determine whether the predicted outcome had occurred. We do not know whether the omission of approximately 300 cases from the determinate forecast set biased the findings.

To explore the possible effects of these two sources of omission, we recomputed *PS* under various conditions. First, we reintroduced the indeterminate forecasts while maintaining the requirement for clear outcome information. This resulted in a set of 1,555 forecasts with a *PS* = .098, only slightly worse than the *PS* = .083 observed for the combined sample of 1,422 forecasts reported in Table 7. Next, we allowed for all outcomes to be represented in the determinate forecast set of 1,721 cases. Like before, clear non-occurrences were coded as 0 and clear occurrences as 1. However, partial non-occurrences were coded as .25 and partial occurrences as .75, while the remaining uncoded outcomes were given an intermediate value of .5. With the inclusion of the additional outcome data, once again *PS* = .098. Finally, all 1,934 forecasts were analyzed by combining these alternative approaches. In this case, *PS* = .106. Thus, even if all the indeterminate forecasts are treated as .5 probabilities and all outcomes are coded regardless of their clarity, the results do not change appreciably. The probability score is still indicative of very good forecasting quality. Thus, we can rule out the methodological issues explored here as the basis of our findings.

While the present study was not initially designed to shed light on the causal determinants of forecasting quality, some of the moderating factors measured nevertheless do help in refuting a number of putative explanations for our findings. For example, it could be argued that analysts' forecasts were of high quality mainly because they dealt with simple forecasting problems that would be hard to get wrong—the proverbial "the world will continue tomorrow" type of forecast. However, analyzing the results by difficulty level does not support that explanation. Calibration was unaffected by the difficulty of the forecast, and, although discrimination was better for the easier forecasts, it was of impressively high quality for the relatively harder forecasts as well.

A related explanation is that forecasting quality was impressive mainly because analysts focused on relatively unimportant judgments. However, the evidence suggests quite the opposite. Forecasts that were coded as more important for policy decision-making were in fact better calibrated and showed better discrimination than those deemed to be of lesser importance. We are not surprised by such findings as we would expect the most important assessments to receive the greatest care from analysts and their directors.

In a similar vein, one might have expected that forecasts from externally tasked reports would be harder for analysts to judge accurately than those in internally tasked reports since the latter affords more control over the subject matter. However, that hypothesis was not supported either. Forecasting skill was in fact better among the set of externally tasked reports than among the set of internally tasked reports (see Table 7). Again, we find these results intuitive since client-requested assessments should elicit a relatively high degree of care from intelligence personnel.

In short, there was little support for a class of cynical explanations that, in different ways, attempt to "explain away" or otherwise diminish the forecasting success observed in this study. The lack of evidence for these accounts is all the more noteworthy given how accessible (and seemingly plausible) they tend to be. Question periods following the authors' presentation of the findings at workshops and conferences have often included comments along these lines.

A different type of putative explanation considered by the lead authors following the analysis of the Phase I forecasts was that the forecasting success observed may have been attributable to the standards and procedures designed to improve analytic integrity that were implemented in the MEA Division. Indeed, this thinking lay behind the expansion of the investigation in Phase II to assess forecasts from other IAS divisions and interdepartmental committees. The results of that investigation, however, did not support the initial hypothesis. Calibration was identical in the two sets of forecasts, and discrimination was somewhat better in the non-MEA Division sample. Earlier, we had urged caution in interpreting these results because indeterminate cases were handled differently in the MEA Division. Forecasts from non-MEA Division sources were excluded if they were issued with indeterminate terms, leading to a more selective set of forecasts for the assessment exercise. However, if we reintroduce indeterminate forecasts, assigning them a probability of .5, then a statistical comparison of MEA and non-MEA Division forecasts reveals no significant difference in probability scores: PS = .094 (SD = .165) for MEA, and PS = .099 (SD = .191) for non-MEA, $t(1553) = -0.47$, $p = .64$. In short, there was no evidence of differential forecasting performance between the MEA Division and other organizational sources sampled in this study.

The absence of an observed forecasting boost for the MEA Division is, in one sense, disappointing and, in another, reassuring. It is disappointing because standards instituted in the MEA Division, such as the use of the standard for uncertainty terms shown in Table 3, were intended to promote and, indeed, improve the quality of intelligence analysis. While the present findings do not rule out the possibility that such standards may have improved the quality of analysis in other ways, nor do they show positive evidence either. From another vantage point, however, these findings are reassuring in that whatever is determining the high degree of forecasting success observed in this study, it appears to be a factor or set of factors, more likely, that are much more pervasive within the strategic intelligence community than the local standards of a single division. Overall, this is good news, although it still leaves open the question of what precisely is accounting for the forecasting success observed.

Although we cannot offer anything approximating a definitive answer to this question, an assessment of the forecasting environment in which strategic intelligence analysts find themselves points to a number of plausible sources. First, there is the level of engagement to consider. Most studies of calibration and confidence rely on university student convenience samples completing tasks that involve answering large numbers of general knowledge questions—as Keren (1987) noted, this is an exceedingly boring task with no personal or professional consequences for the participant.

Studies of expert judgment take engagement up several notches, but many of those still fall far short of the level of engagement characterizing forecasters in this study. In some cases, such as

Tetlock's (2005) landmark study of expert political forecasting, the experts issued forecasts as part of a research study. They knew that their results would be analyzed anonymously and reported in terms of quality only in the aggregate. The present study was quite different. Analysts produced the forecasts as part of their regular professional work. Indeed, the forecasts were not elicited from analysts but extracted from classified intelligence reports in which the analysts were specifically named.

Strategic intelligence forecasting is not only engaging, in the sense of making predictions about something one is interested in; it is also a high-stakes enterprise. The consumers of intelligence are usually one's peers (i.e., other members of the intelligence community) and decision-makers faced with consequential choices for which they are accountable. The realm of such decision-making is itself one that commands attention since it deals directly with matters of national defence, security, and foreign policy. In this respect, studies of strategic intelligence forecasting differs from many other studies of expert judgment, where the stakes are considerably lower—such as bridge playing (Keren, 1987) or even making probability of precipitation forecasts (Murphy & Winkler, 1984).

In some studies, however, the stakes do appear to be high, and yet the level of forecasting or diagnostic accuracy achieved falls far short of that observed in this study. For instance, physicians' hemodynamic assessments of their critically ill patients are overconfident and show only modest discrimination (Dawson et al., 1993). Lawyers' predictions about their case outcomes were also overconfident in one study (Goodman-Delahunty et al., 2010). Moreover, in neither of these two studies did experience improve judgment quality. Dawson et al. (1993) reported that more experienced physicians were more confident in their assessments but no more accurate than less experienced physicians. Goodman-Delahunty et al. (2010) reported significant overconfidence for both junior and senior lawyers.

It is unclear why intelligence analysts' forecasts should be so much better than these experts' probabilistic judgments. While the stakes in the medical and legal domains are not matters of *national* security, they are certainly serious matters for the experts' clients, and it would be hard to imagine that the experts would not feel highly accountable for their judgments.

One potentially important difference between these expert domains is that, whereas the act of judgment (e.g., forecasting or diagnosis) and the act of decision-making are undertaken by distinct groups of experts in the defence and security realm, they are undertaken by the same experts in the medical and legal realms. Physicians who make diagnoses about a patient's medical condition also take what they regard as appropriate action. Lawyers who predict a certain outcome for a case choose a strategy that they deem appropriate in light of those predictions. In other words, these experts make judgments that support their decisions. Their clients are relatively passive, relying not only on the experts' judgments but also, and mainly, on the success of their actions.

Intelligence analysts find themselves in quite a different relationship to the decision-making process and their clients. Intelligence informs defence and security decision-making, but the decisions in such cases are made by commanders or policy officials rather than by intelligence personnel. The intelligence community thus fulfills primarily an advisory rather than an agent-based role. The agents they advise tend to be in positions of considerable power. This may account for some of the differences observed. Advisors are more likely to be blamed for bold forecasts that end up being wrong than for timid forecasts that end up being right. The pattern of under-extremity bias, indicative of timid or under-confident forecasting, observed in virtually every partition of the overall sample of forecasts is consistent with that view. Analytic forecasts

showed exceptional discrimination but tended to be issued with less certainty than appears to be warranted by the outcome data.

Thus, consistent with an advisory role under high accountability conditions to a skeptical audience (Tetlock, Skitka, & Boettger, 1989), analytic forecasts were, generally speaking, cautious and correct. Accountability pressures are associated with reduced overconfidence in predicting personality attributes (Tetlock & Kim, 1987), reduced over-attribution bias (Tetlock, 1985), deeper processing of information (Chaiken, 1980), and better awareness of the informational determinants of one's choices (Cvetkovich, 1978; Hagafors & Brehmer, 1983). However, the conditions under which accountability pressures are manifested have much to do with people's responses to such pressures. Accountability can also promote defensive bolstering of one's position or pre-emptive self-criticism (Tetlock et al., 1989). The arms-length relationship between intelligence and policy would seem to promote a form of accountability in which cognitively complex, "fox-like" thinking is the more probable response by intelligence advisors.

Stated differently, intelligence analysts strive for policy neutrality. They try to avoid having any ideological, political, or bureaucratic stake in any of the judgments that they make. They strive to make the best call based on the available evidence and their knowledge of the situation. Of course, analysts and their directors can probably never completely avoid such pressures and biases—even tradecraft methods meant to free the analyst from mindsets and biases can be shaped by partisan political pressures (Mitchell, 2006)—but policy neutrality is an ideal that is stressed in training and in analysts' day-to-day work. This is not always the case with other experts, especially those associated with organizations that have a particular ideological or political agenda. Those experts may often have a dog in the fight, and that is likely to affect the nature of their judgments.

The forecasts in this study also differ from those examined in other expert studies in another potentially important respect. As noted earlier, the forecasts analyzed in this study are not, strictly speaking, analysts' forecasts, although they are analytic forecasts. The forecasts that appear in the intelligence reports are seldom the result of a single individual arriving at his or her judgment. Rather, in most instances, the analytic forecast is part of an organizational product reflecting the input of the primary analyst, his or her director, and possibly a number of peer analysts. To what extent the input of other intelligence personnel improves forecasting accuracy is unknown. Because others with whom the primary analyst interacts are likely to constructively challenge the analyst, it seems plausible that they would promote the same sort of "fox-like" response to accountability pressures that was noted in the preceding paragraph.

5.1 Experience

Evidence on the effects of experience on judgment accuracy reported in prior studies is mixed. Some studies indicate that experience improves calibration by reducing overconfidence (e.g., Keren, 1987), while other studies show that more experienced experts are more confident than their less experienced counterparts but no more accurate (e.g., Dawson et al., 1993).

In the present study, junior and senior analysts exhibited no difference in calibration. However, there was a marked improvement in the discrimination skill of senior analysts over junior analysts. Whereas junior analysts accounted for approximately 59% of outcome variance with their forecasts, senior analysts accounted for approximately 77%—over a 30% increase in discrimination skill.

Moreover, whereas 60.5% of junior analysts' forecasts were deemed to be of high importance, that figure rose to 84.4% among senior analysts' forecasts, a difference that was statistically

significant, Mann-Whitney U = 137,275.0, $p < .001$. Thus, not only did senior analysts show substantially better discrimination skill than junior analysts, they did so while handling a significantly larger proportion of high-importance forecasts.

While it is encouraging to see the overall high level of forecasting quality in this study, it is also encouraging to see that analytic forecasting also improves with experience. To the best of our knowledge, this is the first study to clearly demonstrate a beneficial effect of experience in the intelligence realm.

5.2 Forecast Time Frame

A notable effect observed in this study was that of forecast time frame on forecast quality. As the temporal window of the forecast increased from under one month to over one year, there was a monotonic increase in forecast quality as measured by discrimination and an opposing decrease in calibration, with the poorest calibration observed in the small sample of forecasts with a time frame of one month or less. The overall effect, however, as measured by skill scores was a decline in forecasting quality as time frame increased.

Goodman-Delahunty et al. (2010) examined the correlation between lawyers' confidence and the number of months between prediction and trial date and found no significant correlation. They did not, however, report on how forecasting quality varied by time frame, and we are not aware of other studies that have examined this issue. At present, it remains unclear why time frame had an effect on forecasting quality in this study.

Indeed, the finding might appear counter-intuitive given that uncertainties are likely to multiply over time. Thus, longer forecast time frames should be coupled with greater uncertainty, which should reduce forecasting accuracy given a constant skill level. However, this would only be the case if one were comparing forecasts of a given temporal window size, such as a 24-hour period, in the near future to the same window in a more distant future. In contrast, the time frame variable coded in the present study conflates distance into the future with window size, such that the two are directly proportional. That is, a short time frame forecast might be of the form "X is likely to increase sharply within the next month," while a comparable longer time frame forecast might be "X is likely to sharply increase within the next year." Clearly, the longer time frame affords more opportunities for the predicted event to occur, which should help forecast accuracy in general and discrimination in particular. Future research might examine the effect of time frame under more controlled conditions that independently manipulate temporal window size and distance into the future.

5.3 Implications

As noted earlier, this study is unique. It is the first field study ever conducted to examine the quality of strategic intelligence forecasts over a large set of assessments from real intelligence reports using quantitative scoring rules that are commonly applied in other areas of expert judgment monitoring.

Our investigation was not primarily theoretical. We did not begin hoping to test a series of well-formed predictions derived from one or more basic theories of human judgment. Rather, the investigation, as noted in the second chapter of this report, began solely as a managerial exercise in quality control monitoring and subsequently evolved into a field study. Since the quality of intelligence forecasts had not been previously scored in any systematic manner, it was "simply"

of interest to discover how good intelligence forecasts were. This question, quite remarkably, had not been answered before—and not only in Canada.

5.3.1 Generative Impact

The first author (Mandel) briefed the initial results from the first phase of this study at the US National Academies' National Research Council (NRC) Committee on Behavioral and Social Science Research to Improve Intelligence Analysis for National Security (Mandel 2009c), a committee tasked with making recommendations for intelligence reform via the social and behavioural sciences to its client, the Office of the Director of National Intelligence (ODNI). In its final report, the Committee wrote:

> We recognize that there has historically been resistance to numeric probability estimates from analysts who believe they imply artificial precision. However, as discussed in Chapter 2, the scientific evidence, including Canada's real-world success with numeric probabilities in intelligence analysis (Mandel, 2009), suggest that, with proper training and feedback, such judgments could substantially improve analytic products and customer understanding of them. Proper incentives seek to encourage learning, not to determine culpability. They reward positive performance and cultivate the natural desire to do well, a desire that is especially prevalent in the IC [intelligence community]. In addition, numeric probabilities allow feedback that is essential to learning. Proper incentives discourage both overconfidence (intended perhaps to carry an argument) and underconfidence (intended perhaps to avoid responsibility). They encourage good calibration: being as confident as one's understanding warrants. Thus, DNI [Director of National Intelligence] must ensure that numeric probabilities are implemented in a constructive way, using them for useful feedback, not destructive criticism.
>
> Recommendation 2
>
> The Director of National Intelligence should ensure that the intelligence community adopts scientifically validated analytical methods and subjects all of its methods to performance evaluation. To that end, each analytical product should report, in a standardized format, the elements necessary for such evaluation, including its analytical method, domain, conclusions, analysts' background, and the collaboration that produced it. Analyses must include quantitative judgments of the probability and uncertainty of the events that they forecast. These reports should be archived in a database that is routinely used to promote institutional learning and individual training and as input to the Director of National Intelligence's ongoing review efforts of analytic shortfalls and plans to address them.
>
> Immediate Actions
>
> 1. Institutionalize an "Analytical Olympics," with analysts and analytical methods competing to provide the best calibrated probabilities (i.e., showing appropriate levels of confidence) in assessments and predictions made for well-specified outcomes that have occurred or will occur in the near future.
>
> 2. Begin assessing how well-calibrated individual analysts are, using the results as personal feedback that will allow analysts to improve their own performance and

IC to learn how this performance is related to workforce factors, such as personal capabilities, training, and incentives.

3. Create a research program that reviews current and historic assessments, looking for correlates of accuracy and calibration, considers properties such as method used, collaboration process, classification level, substantive domain, and team composition. (National Research Council, 2010, pp. 86–87)

The impact of not only our research effort but of the longer-term effort of the second author (Barnes) to require MEA Division analysts to use numeric probabilities in their intelligence forecasts is evident in the NRC report. Each of the three immediate actions calls for quantitative scoring of forecast quality and envisions a variety of functions that such activity could serve.

The US IC has already made significant progress in realizing Action 1. Soon after the aforementioned NRC meeting, Jason Matheny, a Program Manager at ODNI's Intelligence Advanced Research Projects Activity (IARPA), contacted the first author to discuss how IARPA might support an S&T program in the area of improving forecasting accuracy. That discussion, and the present research that prompted it (Matheny, personal communication, December 4, 2012), contributed to the development of the Aggregative Contingent Estimation (ACE) Project, a multi-million dollar project that has funded multiple interdisciplinary teams of scientists that are competing to find ways of improving forecasting accuracy through superior methods of sampling, forecast elicitation, and aggregation of forecasts obtained in prediction markets, which involve some form of incentivized crowd sourcing for predictions on predetermined topics (see below). (The first author is a co-investigator on one of the teams led by Charles Twardy and Kathryn Laskey at the Center of Excellence in Command, Control, Communications, Computing and Intelligence at George Mason University.)

The US IC has also made progress on Action 2. For instance, ODNI has established a classified prediction market for US intelligence analysts. As with ACE, the IC prediction market focuses on questions that will eventually resolve into clear answers. Thus, much of the subjectivity involved in coding outcomes and inferring probabilities in the present study is removed in these endeavours. On the other hand, responses to a prediction market do not constitute finished intelligence. What such exercises gain in internal validity, they also lose (at least to some degree) in external validity. Nevertheless, we view these US developments very positively and see our mutual efforts as complementary as well as mutually beneficial in light of our collaborative relationship with many of the individuals leading the US activities just noted.

5.3.2 Outcome-oriented Accountability Systems

The bulk of the effort to promote analytical rigour and account for the quality of analytical products has been process oriented (Tetlock & Mellers, 2011b). For instance, in the US IC, ODNI Intelligence Community Directive (ICD) 203 on "Analytic Standards" (ODNI, 2007) recommends that intelligence analysis should

(a) properly describe the quality and reliability of underlying sources;

(b) properly caveat and express uncertainties or confidence in analytic judgments;

(c) properly distinguish between underlying intelligence and analysts' assumptions and judgments;

(d) incorporate alternative analysis where appropriate;

(e) demonstrate relevance to US national security;

(f) use logical argumentation;

(g) exhibit consistency in analysis over time, or highlight changes and explain the rationale; and

(h) make accurate judgments and assessments.

Of these eight standards for promoting rigour, only the last one lends itself to an outcome-oriented evaluation. The other standards would have to be assessed in terms of the extent to which current practices adopt rigorous processes aimed at promoting the desired qualities that ICD 203 identifies. However, even the accuracy aim outlined in ICD 203 is largely couched in terms of process recommendations:

> Analytic elements should apply expertise and logic to make the most accurate judgments and assessments possible given the information available to the analytic element and known information gaps. Where products are estimative, the analysis should anticipate and correctly characterize the impact and significance of key factors affecting outcomes or situations. (ODNI, 2007, p. 4)

A single sentence acknowledges the difficulty of outcome accountability in this regard: "Accuracy is sometimes difficult to establish and can only be evaluated retrospectively if necessary information is collected and available" (ODNI, 2007, p. 4).

As noted earlier, the present fieldwork is unique in that it is the first study to evaluate systematically and comprehensively the forecasting accuracy of intelligence products over several years. Although the study has been laborious, we have shown that it can be done with reasonable effort and a modest budget. It is possible for the intelligence community to apply the same sorts of scoring rules that have been employed in other domains of expert judgment, and it is our hope that our approach might serve as a template or model for others who wish to adopt an outcome-oriented approach to assessing forecast accuracy in intelligence. We believe that the greatest impediments to implementing an outcome-oriented system for ongoing assessment of forecasting quality are motivational and knowledge-based in nature. Simply put, too few directors, at present, see the need for objectively scoring the accuracy of intelligence forecasts, and fewer still are aware of the methods required to implement such a system. Exceptions to the rule are likely to retire (like the second author) before widespread institutional change takes hold. It may take a push for change from top levels of management or even elected (or appointed) officials before an outcome-based accountability system is implemented. It remains to be seen whether the US IC, for instance, will follow through on the NRC recommendations to systematically track the accuracy and calibration of their forecasts.

5.3.3 Performance Feedback and Adaptive Learning

While our study serves as an example of how to implement an outcome-oriented approach to monitoring forecast quality, it also suggests a model for adaptive learning through performance feedback. That is, the type of results gathered in the present study could be provided to analysts at an individual level, enabling them to see how well they have been forecasting over time. Rieber (2004) has advocated just such an approach, recommending that intelligence analysts be given calibration feedback on their performance. Moreover, there is some evidence to suggest that performance-based feedback can improve judgment quality (e.g., Alpert & Raiffa, 1982; Benson & Önkal, 1992; Bolger & Önkal-Atay, 2004). For instance, Sharp et al., (1988) found that

students given detailed feedback regarding their confidence and accuracy rates showed improvements in discrimination in their subsequent judgments, whereas students given no feedback showed no change. The performance feedback, however, did not seem to improve calibration. Conversely, Stone and Opel (2000) found that performance feedback improved calibration but did not influence discrimination.

Given that both calibration and discrimination were already very good in the present study, it would be of interest to see whether it would be possible to improve upon prior performance. Since we did not examine forecast quality at the individual level, it is unclear how much variance in forecasting quality is accounted for by differences between analysts. The difference in discrimination between senior and junior analysts, however, does suggest that variations in at least some aspects of quality exist *between* analysts. Indeed, one might hypothesize that junior analysts on average are not only poorer in discrimination than senior analysts but that they are also more variable. Perhaps the best junior analysts are as good as the best senior analysts, but there may also be a subset of much poorer-performing junior analysts.

To provide an initial test of this hypothesis, we computed Levene's test for the equality of variances on *PS,* which revealed that the variability among junior analysts (SD = 0.222) was in fact significantly greater than the variability among senior analysts (SD = 0.159), F = 39.9, $p <$ .001. This initial test, however, is only suggestive because the sample size in the senior sample is also greater. Thus, the lower variability among senior analysts may simply reflect a more stable estimate. To assess the impact of unequal sample sizes on the Levene test, three random sub-samples of senior analysts—each equal in size to the smaller junior analyst sample—were drawn from the senior analyst sample. In each of these three cases, the same Levene test was calculated, and in all three cases the test was significant at $p <$ .001. Thus, the inequality in *PS* variance between junior and senior analysts does not appear to be due to a mere sample size inequality. Junior analysts' forecasting performance does indeed appear to be more variable than senior analysts' performance.

5.3.4 Post-Forecast Debiasing

Although this study revealed a high degree of forecasting quality, a systematic pattern of miscalibration was also evident. As noted earlier, that pattern revealed a tendency for analytic forecasts to be communicated with less certainty than warranted. This may reflect an organizational disposition towards caution or even understatement. Such a tendency may be better than the alternative: overconfident pronouncements that end up being wrong. Nevertheless, there is a cost associated with overly timid forecasts that couch prediction in unnecessary degrees of uncertainty. Decision-makers often want clear, unambiguous answers. Unnecessary uncertainty expressed in forecasts may water down the indicative value of communications, making them less salient and possibly less influential in subsequent decision-making. To the extent that miscalibration is unsystematic, there is not much that can be done to correct it. However, such opportunities for correction do exist when there is a systematic element in observed judgment error. Aside from using that knowledge to help analysts improve their forecasting through adaptive learning, steps can be taken to adjust forecasts after the fact in ways that exploit the systematicity in judgment error.

Consider, for instance, the process followed in the MEA Division in which analysts assigned probabilities to forecasts, yet those probabilities were only indirectly communicated in the intelligence reports through verbal expressions of uncertainty that may have captured more or less well the assigned numeric probability. In this case, there is an opportunity to remap probabilities to minimize miscalibration. As Figure 3 showed, there were roughly four levels of probability differentiated by analytic forecasts. Probabilities of 0/10, 1/10, and 2.5/10 had a relative

frequency of about 5%; probabilities of 4/10 and 5/10 had a relative frequency of about 25%; probabilities of 6/10 and 7.5/10 had a relative frequency of about 85%; and probabilities of 9/10 and 10/10 had a relative frequency of about 95%. Based on this information, one could simply remap the original probabilities such that values of 0, 1, and 2.5 are remapped to .5 (out of 10); values of 4 and 5 are remapped to 2.5; values of 6 and 7.5 are remapped to 8.5; and values of 9 and 10 are remapped to 9.5. This simple remapping process would not affect discrimination at all, but it would substantially improve calibration. Indeed, Figure 21 shows the reliability diagram for the MEA Division forecasts after being remapped as just described. The calibration index improves from a value of .014 (unmapped, as reported in Table 5) to a value of .0002.

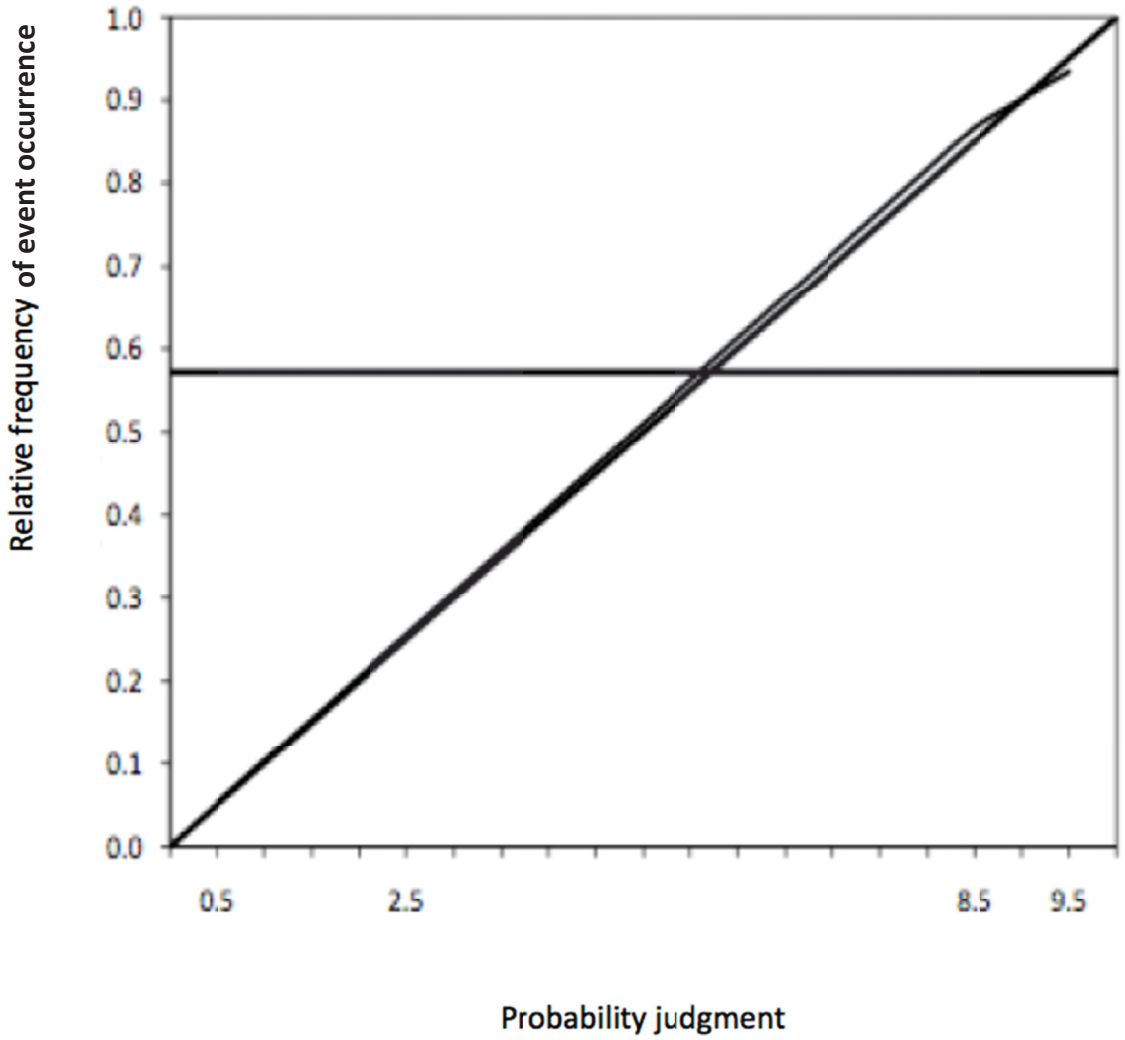

Figure 21: Reliability Diagram showing Remapped MEA Division Forecasts

Of course, the level of improvement observed is to be expected given that the remapping capitalizes on what was already known from the data. If the systematicity in those data is not robust, then the value of such optimization procedures would suffer on cross-validation trials or future applications. There is, however, good reason to believe that post-forecast debiasing through optimization would be effective because the pattern of bias—namely, underextremity with some overconfidence at the extremes of 0 and 1—exploited by the current remapping procedure was robust across levels of the moderator variables explored in this study as well as in the forecast data based on inferred probabilities.

This optimization procedure could also be implemented at finer levels of forecast aggregation, provided there are sufficient data for reliable characterization of bias patterns at these levels. For instance, these procedures could be applied to the forecasts of individual analysts. One might imagine that, while IAS forecasts overall exhibit a pattern of underextremity bias, some IAS analysts might show different, even opposing patterns. If analysts' biases were well known, post-forecast optimization rules could be applied to their forecasts, thus proving a more tailored optimization solution. More generally, as more reliable estimates of error and bias for subpopulations of forecast become known, more can be done to improve future judgment quality through judgment-support technologies.

Such knowledge can also benefit the aggregation of forecasts for decision-makers or other intelligence fusers. Linear opinion pools (Clemen & Winkler, 1999; Winkler & Clemen, 2004)—which take an unweighted average of the forecasts on a topic generated by multiple, independent forecasters—can improve forecasting accuracy (e.g., Surowiecki, 2004) as can averaging multiple judgments on a topic made by the same forecaster (e.g., Vul & Pashler, 2008). Part of the reason for such improvements resulting from simple aggregation methods such as unweighted averaging is that unsystematic error in judgment tends to get cancelled out, leaving more indicative signals to be communicated to forecast receivers. However, aggregation methods can also capitalize on systematicity in error. For instance, some methods assign weight to forecasters in inverse proportion to their degree of exhibited coherence in judgment (Cooke & Goossens, 2008; Karvetski, Olson, Mandel, & Twardy, 2013; Wang, Kulkarni, Poor, & Osherson, 2011). Likewise, it is not difficult to imagine how a decision-maker might pool different analysts' forecasts, assigning greater weight to those analysts having exhibited better forecasting skill in the past. Such processes could be automated with decision-support technologies so that decision-makers do not have to "manually" aggregate the forecasts they receive. While such endeavors might appear as no more than intellectual flights of fancy at present, efforts are in fact already underway to design such systems for intelligence consumers and providers. IARPA's ACE competition, noted earlier, is a good example of how the US IC is investing now in developing "crowdsourcing" technologies for future strategic intelligence requirements.

6 Concluding Remarks

The present study represents a unique experiment in the intelligence community, one in which intelligence judgments of a predictive nature were systematically extracted from finished products and coded in terms of multiple attributes. Outcome data were diligently collected and used to score the forecasts in terms of a number of quantitative indices of forecast quality that have been used in other areas to monitor the quality of expert judgment. The forecasts tracked in this study span roughly a half-decade and include the bulk of forecasts produced by IAS's MEA Division over that period. The act of studying forecast quality itself led to refinements in the process for doing so, and the undertaking has served as a unique learning experience for the authors and a generative force for the application of forecasting science to the realm of intelligence analysis.

In closing, this study has revealed a high degree of forecasting quality in Canadian strategic intelligence assessments, both in terms of the calibration and discrimination of analytic forecasts. It also shed light on the degree to which various moderating factors—such as the analyst's experience, the forecast's difficulty and importance, etc.—influenced measures of forecast quality. And it provided a characterization of the nature of bias present in this sample of organizational forecasts. Taken together, the findings suggest ways in which the research approach adopted in this study could be used to develop outcome-based accountability systems, adaptive learning systems, and optimization procedures to support informed and effective decision-making.

References

Alpert, M., & Raiffa, H. (1982). A progress report on the training of probability assessors. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 294–305). Cambridge: Cambridge University Press.

Arkes, H.R., & Kajdasz, J. (2011). Intuitive theories of behavior. In B. Fischhoff & C. Chauvin, (Eds.), *Intelligence analysis: Behavioral and social scientific foundations* (pp. 143–168). Washington, DC: National Academies Press.

Åstebro, T., & Koehler, D.J. (2007). Calibration accuracy of a judgmental process that predicts the commercial success of new product ideas. *Journal of Behavioral Decision Making*, *20*, 381–403.

Benson, P.G., & Önkal, D. (1992). The effects of feedback and training on the performance of probability forecasters. *International Journal of Forecasting*, *8*, 559–573.

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

Bolger, F., & Önkal-Atay, D. (2004). The effects of feedback on judgmental interval predictions. *International Journal of Forecasting, 20*(1), 29–39.

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

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

Clemen, R.T., & Winkler, R.L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis, 19*(2), 187–203.

Cooke, R.M., & Goossens, L.L.H.J. (2008). TU Delft expert judgment data base. *Reliability Engineering and System Safety*, *93*, 657–674.

Cvetkovitch, G. (1978). Cognitive accomodation, language, and social responsibility. *Social Psychology Quarterly, 41*, 149–155.

Dawson, N.V., Connors, A.F., Speroff, T., Kemka, A., Shaw, P., & Arkes, H.R. (1993). Hemodynamic assessment in managing the critically ill: Is physician confidence warranted? *Medical Decision Making, 13*(3), 258–266.

Erev, I., Wallsten, T.S., & Budescu, D.V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, *101*(3), 519–527.

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

Gardner, D. (2010). *Future babble.* New York, NY: Penguin.

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge, UK: Cambridge University Press.

Goodman-Delahunty, J., Granhag, P.A., Hartwig, M., & Loftus, E.F. (2010). Insightful or wishful: Lawyers' ability to predict case outcomes. *Psychology, Public Policy & Law*, *16*(2), 133–157.

Greenwald, A.G. (2012). There is nothing so theoretical as a good method. *Perspectives on Psychological Science, 7*(2), 99–108.

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

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

Karvetski, C.W., Olson, K.C., Mandel, D.R., & Twardy, C.R. (2013). *Probabilistic coherence weighting for optimizing expert forecasts*. Manuscript submitted for publication.

Kent, S. (1964). Words of estimative probability. Reprinted in D. P. Steury (Ed.), *Sherman Kent and the Board of National Estimates*. Washington, D.C.: Central Intelligence Agency, Center for the Study of Intelligence. Retrieved from https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/books-and-monographs/sherman-kent-and-the-board-of-national-estimates-collected-essays/6words.html

Keren, G. (1987). Facing uncertainty in the game of bridge: A calibration study. *Organizational Behavior and Human Decision Processes, 39*, 98–114.

Kesselman, R.F. (2008). *Verbal probability expressions in national intelligence estimates: A comprehensive analysis of trends from the fifties through post 9/11* (Master's thesis). Mercyhurst College, Erie, PA.

Koehler, D.J., Brenner, L., & Griffin, D. (2002). The calibration of expert judgment: Heuristics and biases beyond the laboratory. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 686–715). Cambridge, UK: Cambridge University Press.

Liberman, V., & Tversky, A. (1993). On the evaluation of probability judgments: Calibration, resolution, and monotonicity. *Psychological Bulletin*, *114*(1), 162–173.

Lichtenstein, S, Fischhoff, B., & Phillips, R (1982). Calibration of probabilities: The state of the art to 1980. D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). New York: Cambridge University Press.

Lin, S-W., & Bier, V.M. (2008). A study of expert overconfidence. *Reliability Engineering and System Safety*, *93*, 711–721.

Mandel, D.R. (2008a). *A calibration study of an intelligence assessment division*. Paper presented at the Inter-University Seminar on Armed Forces and Society Canada Region Conference, Kingston, Ontario.

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

Mandel, D.R. (2009a). *A calibration study of an intelligence assessment division*. Paper presented at the 22nd Research Conference on Subjective Probability Utility Decision Making (SPUDM22), Rovereto, Italy.

Mandel, D.R. (2009b). *Canadian perspectives: 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 Roundtable "What Can the Cognitive and Behavioural Sciences Contribute to Intelligence Analysis? Towards a Collaborative Agenda for the Future," Meech Lake, Quebec.

Mandel, D.R. (2009c). *Canadian perspectives: Applied behavioral science in support of intelligence analysis*. Invited paper presented at the Public Workshop of the National Academy of Sciences Committee on Behavioral and Social Science Research to Improve Intelligence Analysis for National Security, Washington, D.C.

Mandel, D.R. (2011a). *Understanding and augmenting human analytic capabilities*. Paper presented at the Intelligence Way Ahead Symposium, Ottawa, Ontario.

Mandel, D.R. (2011b). *Validating the quality of predictive intelligence assessments: Towards a proactive outcome-based accountability system.* Paper presented at the Canadian Association of Professional Intelligence Analysts Spring Conference on Enabling Technologies for Intelligence Analysts, Ottawa, Ontario.

McClelland, G.H. (2011). Use of signal detection theory as a tool for enhancing performance and evaluating tradecraft in intelligence analysis. In B. Fischhoff & C. Chauvin, (Eds.), *Intelligence analysis: Behavioral and social scientific foundations* (pp. 83–100)*.* Washington, DC: National Academies Press.

Mitchell, G.R. (2006). Team B intelligence coups. *Quarterly Journal of Speech, 92*(2)*,* 144–173.

Moore, D.A., & Healy, P.J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502–517.

Murphy, A.H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology, 12*(4), 595–600.

Murphy, A.H., & Winkler, R.L. (1984). Probability forecasting in meteorology. *Journal of the American Statistical Association, 79*(387), 489–500.

National Research Council. (2010). *Planning committee on field evaluation of behavioral and cognitive sciences-based methods and tools for intelligence and counterintelligence.* Washington, DC: National Academies Press.

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

Office of the Director of National Intelligence (ODNI). (2007). *Analytic standards* [Intelligence Community Directive (ICD) 203]. Retrieved from https://www.fas.org/irp/dni/icd/icd-203.pdf

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

Sharp, G.L., Cutler, B.L., & Penrod, S.D. (1988). Performance feedback improves the resolution

of confidence judgments. *Organizational Behavior and Human Decision Processes, 42,* 271–283.

Stone, E.R., & Opel, R.B. (2000). Training to improve calibration and discrimination: The effect of performance and environmental feedback. *Organizational Behavior and Human Decision Processes, 83*(2), 282–309.

Surowiecki, J. (2005). *The wisdom of crowds.* New York, NY: Double Day.

Tetlock, P.E. (1985). Accountability: A social check on the fundamental attribution error. *Social Psychology Quarterly, 48,* 227–236.

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

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

Tetlock, P.E., & Mellers, B.A. (2011a). Intelligent management of intelligence agencies: Beyond accountability ping-pong. *American Psychologist, 66* (6), 542–554.

Tetlock, P.E., & Mellers, B.A. (2011b). Structuring accountability systems in organizations: Key trade-offs and critical unknowns. In B. Fischhoff & C. Chauvin, (Eds.), *Intelligence analysis: Behavioral and social scientific foundations* (pp. 249–270). Washington, DC: National Academies Press.

Tetlock, P.E., Skitka, L., & Boettger, R. (1989). Social and cognitive strategies for coping with accountability: Conformity, complexity, and bolstering. *Journal of Personality and Social Psychology, 57*, 632–640.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science,* 19*(7)*, 645–647.

Wang, G., Kulkarni, S.R., Poor, H.V., & Osherson, D.N. (2011). Aggregating large sets of probabilistic forecasts by weighted coherent adjustment. *Decision Analysis*, *8*, 128–144.

Winkler, R. (1994). Evaluating probabilities: Asymmetric scoring rules. *Management Science, 40,* 1395–1405.

Winkler, R.L., & Clemen, R.T. (2004). Multiple experts vs. multiple methods: Combining correlation assessments. *Decision Analysis, 1*(3), 167–176.

Yaniv, I., Yates, J.F., & Smith, J.E.K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin*, *110*(3), 611–617.

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

This page intentionally left blank.

List of acronyms

ACH	Analysis of Competing Hypotheses
ACE	Aggregative Contingent Estimation
CDI	Chief of Defence Intelligence
CI	Calibration index
DI	Discrimination index
DMIAC	Deputy Minister Intelligence Assessment Committee
DND	Department of National Defence
DNI	Director of National Intelligence
DRDC	Defence Research & Development Canada
DRDKIM	Director Research and Development Knowledge and Information Management
IACC	Intelligence Assessment Coordinating Committee
IARPA	Intelligence Advanced Research Projects Activity
IAS	International Assessment Staff/Intelligence Assessment Secretariat
IC	Intelligence community
ICD	Intelligence Community Directive
MEA	Middle East & Africa
MOU	Memorandum of Understanding
NRC	National Research Council (US)
PCO	Privy Council Office
PS	Probability score
ODNI	Office of the Director of National Intelligence
R&D	Research & Development
S&T	Science & Technology
US	United States
VI	Variability index

DOCUMENT CONTROL DATA

(Security classification of title, body of abstract and indexing annotation must be entered when the overall document is classified)

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

Defence R&D Canada – Toronto
1133 Sheppard Avenue West
Toronto, Ontario M3M 3B9

2. SECURITY CLASSIFICATION (Overall security classification of the document including special warning terms if applicable.)

UNCLASSIFIED
(NON-CONTROLLED GOODS)
DMC A
Review: GCEC APRIL 2011

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

A Quantitative Assessment of the Quality of Strategic Intelligence Forecasts

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

Mandel, D. R.; Barnes, A.; Richards, K.

5. DATE OF PUBLICATION (Month and year of publication of document.)	6a. NO. OF PAGES (Total containing information, including Annexes, Appendices, etc.)	6b. NO. OF REFS (Total cited in document.)
March 2014	78	56

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

Technical Report

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

Defence R&D Canada – Toronto
1133 Sheppard Avenue West
Toronto, Ontario M3M 3B9

9a. PROJECT OR GRANT NO. (If appropriate, the applicable research and development project or grant number under which the document was written. Please specify whether project or grant.)	9b. CONTRACT NO. (If appropriate, the applicable number under which the document was written.)
15dm02	

10a. ORIGINATOR'S DOCUMENT NUMBER (The official document number by which the document is identified by the originating activity. This number must be unique to this document.)	10b. OTHER DOCUMENT NO(s). (Any other numbers which may be assigned this document either by the originator or by the sponsor.)
DRDC Toronto TR 2013-036	

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

Unlimited

12. DOCUMENT ANNOUNCEMENT (Any limitation to the bibliographic announcement of this document. This will normally correspond to the Document Availability (11). However, where further distribution (beyond the audience specified in (11) is possible, a wider announcement audience may be selected.))

Unlimited

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

This report describes a field study of the quality of probabilistic forecasts made in Canadian strategic intelligence reports. The researchers isolated a set of 1,422 probabilistic forecasts from intelligence memoranda and interdepartmental committee reports for which outcome information about the forecasted events was available. These data were used to study forecast quality measures, including calibration and discrimination indices, commonly employed in other areas of expert judgment monitoring research (e.g., meteorology or medical diagnosis). Predictions were further categorized in terms of other variables, such as the organizational source, forecast difficulty, and forecast importance. Overall, the findings reveal a high degree of forecasting quality. This was evident in terms of calibration, which measures the concordance between probability levels assigned to forecasted outcomes and the relative frequency of observed outcomes within that assigned category. It was also evident in terms of adjusted normalized discrimination, which measures the proportion of outcome variance explained by analysts' forecasts. The main source of bias detected in analytic forecasts was underconfidence: Analysts often rendered forecasts with greater degrees of uncertainty than were warranted. Implications for developing outcome-oriented accountability systems, adaptive learning systems, and forecast optimization procedures to support effective decision-making are discussed.

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

forecasting, probability judgment, strategic intelligence, calibration, discrimination, skill.

Defence R&D Canada

Canada's Leader in Defence
and National Security
Science and Technology

R & D pour la défense Canada

Chef de file au Canada en matière
de science et de technologie pour
la défense et la sécurité nationale

DEFENCE R&D DÉFENSE

www.drdc-rddc.gc.ca