Improving Forecasting for Foreign Policy*

Identifying drivers of forecasting accuracy in a forecasting tournament

Alexander Sacharow†

April 28, 2017

Abstract

In this paper, I explore the drivers of forecasting accuracy for geopolitical events with the help of a forecasting tournament. The paper analyses the responses of the participants in order to test and evaluate several explanations of successful forecasting. More specifically, it looks at (1) measurable characteristics of forecasters, (2) the decision environment in which a forecast is made and (3) minimal interventions aiming at improved forecasting judgements. My findings are that intelligence is a good indicator of forecasting success reinforcing prior findings on geopolitical forecasting. There is also some evidence that moral judgment competency is related to forecasting accurarcy, but further research is needed. Regarding the decision context, the forecasting tournament showed that more forecasting time is related to more accuracy, but with diminishing returns. Finally, the forecasting competition indicated that small interventions in form of analytical guides cannot improve forecasting judgements. These insights can be used to compute improved crowd forecasts and inform policy makers engaged in forecasting. The results were derived from a security policy forecasting tournament which took place from February to April 2017 and which had more than 200 participants, comprised out of university students with a strong interest in the field, paid online respondents and voluntary online users.

Contents

*A summary of the forecasting tournament and its results are available on the project website. Additional background information and methodological details can be access via the online appendix. The r scripts are available on the author's GitHub account. For the raw data please contact the author.

†Hertie School of Governance, corresponding address: a.sacharow@mpp.hertie-school.org

1 Introduction

Foreign policy makers, just like other decision makers, have to constantly think about the future and how it can unfold. They need to have an idea of possible outcomes and their likelihood in order to make their decisions. For this, they rely mainly on explicit or implicit forecasting judgements, either by themselves, their advisors or some outsiders. However, forecasting geopolitical trajectories in an uncertain world has proven to be a challenge. Too often forecasts are flawed and therefore unreliable for decision makers. In order to improve future-oriented foreign policy making, a better understanding of successful forecasting is crucial.

There are several ways how this can happen. First, a better understanding leads to better forecasts which are a prerequisite for pro-active policy. Foreign policy is often described as dominated by reactive policy making. In order to break this pattern, more reliable methods for forecasting possible futures are essential. Second, it will draw the attention to ill-conceived assumptions underlying today's decision-making and thereby offers a chance to address them appropriately. Third, a better understanding of forecasting is necessary to transform institutions tasked with forecasting. Knowledge about individual differences between forecasters can be used to staff and structure such agencies. A better understanding of the decision environment will allow changing this environment to make it more suitable for forecasting. Likewise, tested decision aids can be used to counter common decision-making flaws made by forecasters. This research focuses primarily on the third point, but indirectly it does also contribute to the other points.

In this paper, a forecasting tournament is used to identify drivers behind accurate forecasting. The tournament is used both to test prior findings in the literature and to deepen the

knowledge on forecasting accuracy by testing new hypotheses.

Compared to other forecasting modes, a tournament has the advantage to force forecasters to make precise predictions and it evaluates these predictions based on their track record. This reduces the level of ambiguity in forecasts, which is common for many other forms of forecasting, and measures the quality of forecasts against what they actually attempt to do.

However, it does also highlight the limitations of this research: A forecasting tournament defines successful forecasting in terms of accuracy. Participants have to specify the likelihood of a particular event in a given time frame in probability and they perform well if they choose probabilities close to the truth. Other indicators for successful forecasting are sidelined, e.g. identifying relevant possible future events or specifying the impact of certain future events.

This paper is structured as follows: First, the literature on forecasting and in particular forecasting competitions is reviewed and critically discussed. In the second section the hypotheses of this research and their theoretical as well as empirical background are presented. Third, the set-up of the research design is explained and some crucial design choices are reviewed. Fourth, the results of the research are presented. Fifth, the results are used to aggregate the individual forecasts. Finally, the research in general is discussed and some policy recommendations are drawn.

2 Literature Critique

In order to forecast one has to make a basic assumption: Forecasting is possible at all. Not everyone, however, agrees to this. Forecasting sceptics emphasize the fundamental uncertainty of the future and assume successful predictions about world politics are ultimately grounded in luck (Almond and Genco 1977; Beyerchen 1992; Taleb 2007). They don't claim that forecasting in general is impossible, but in their view the fundamental problems about foreseeing the future are particular strong in the field of international politics. And they have a point, as foreign policy has many conditions which have proven to be unfavorable for forecasting: The environment is dynamic, most events are essentially unique, feedback on forecasts has long delays, there is a lack of empirical tested decision aids and a strong reliance on subjective judgments (Shanteau 1992). There are even indications that the quality of judgements gets worse with professionalization as intelligence experts specialized in forecasting seem to exhibit even more decision-making

biases than college students (Reyna et al. 2014).[1]

But others scholars are more optimistic about the prospects of forecasting in the field. In their view, forecasting foreign policy is still in its infancy which is partially attributed to the dominant role explanation enjoyed in the past among international relation scholars (Michael D. Ward 2016). This is, however, gradually changing as a result of more forecasting-oriented research. On the individual level, the research has shown grave differences between individual forecasters, making forecasting not only a question of how to forecast but also of who is forecasting (P. Tetlock 2005; Bueno de Mesquita 2009; B. Mellers et al. 2015) and that forecasting can be further improved by appropriated training (B. Mellers et al. 2014). There has also been a rise in the number of methods available and the data used to generate more sophisticated forecasts for world politics (Dhami et al. 2015; Michael D. Ward 2016).

This research is based on the presumption that forecasting is possible, but to a limited extent. Forecasting will never produce a fully certain prediction of the future and some events and aspects will remain beyond the forecastable. However, there are aspects which can be improved and the level of uncertainty can be reduced in a systematic and reliable manner.

The question is how to improve foreign policy forecasting in a meaningful way and what method to use for it. Generally, a wide variety of approaches is available: Starting from the simplest and most common one: Intuitive predictions. These are statements people make out of their head without recourse to a systematic methodology. The good thing about these predictions is that they can be applied to most topics at almost no cost. Unfortunately, they have a record of being inaccurate. Take for example affective predictions which rarely match the actual experience (Wilson and Gilbert 2005; Schkade and Kahneman 1998) or probability judgments which have been shown to be susceptible to biases (D. Kahneman and Tversky 1974). The most simple statistic models have shown to outperform intuitive predictions in various domains like university admission or parole violations (R. Dawes, Faust, and Meehl 1989; Swets, Dawes, and Monahan 2000). More recent research has demonstrated that expert prediction exhibit the same problems. In the political sphere, by tracking expert statements for more than 20 years P. Tetlock (2005) has argued expert predictions are often as accurate as a "dart-throwing chimpanzee". Similar results can be found with experts in other fields, e.g. climate science (Green and Armstrong 2007). One of the reasons for this discrepancy is that forecasting and explanation require different set of skills and there is no reason that experts combine both of them. Another reason

[1] Which has been attributed to bad habits developed in their working environment.

is that experts tend not to apply the same rigidity to forecasting statements when asked about it than to their, often written, work. This does not mean expert opinions should be disregarded, but for forecasting purposes they should be treated with care.

One approach to solve the problems of inaccurate forecasting is to rely more on quantitative methods, as it is common in other disciplines like meteorology or economics. Statistical forecasting has been discussed for long in the sphere of international relations (Choucri 1974) and the field is clearly on the rise (Michael D. Ward 2016). Well known applications are election forecasting models (e.g. Lewis-Beck 2005; Norpoth and Gschwend 2010) or the work of the Political Instability Task Force (Goldstone et al. 2010), which led to the Integrated Crisis Early Warning System (ICEWS) Project. The forecasts are generally based on measurable input variables ranging from economic, media to political indicators. Using past data the models are calibrated and then used for extrapolating into the future.

But there are also severe limitations to this approach. First, statistical models are largely limited to quantifiable events which can be grouped by their similarity. Hence, possible forecasts are, for example, about the outbreak and scale of violence or protests. Many events in foreign policy, however, are at least to some degree unique. Take for instance a court ruling of the International Court of Justice on a specific matter. It is hard to impossible to build a statistical model for such types of events. Hence, there are often no statistical models available to forecast relevant events or important information has to be neglected in order to make events predictable with quantitative models. Second, many statistical models have focused on finding significant relationships instead of useful predictive indicators (M. D. Ward, Greenhill, and Bakke 2010). As a result, they do not produce precise predictions and therefore lack the external validity to be useful for actual forecasting. A well-known example is the Flu forecast developed by Google which could not produce accurate predictions after its initial introduction (Lazer et al. 2014). Moreover, in many cases the necessary input data for quantitative models is scarce, not available or too expensive to gather. This limits the usefulness of quantitative models for policy makers. However, if there are statistical forecasting models with a good track record available, using them is surely a promising approach. The focus of this research is, however, more on geopolitical events for which quantitative models so far haven't been able to produce useful forecasts.

Prediction markets are another approach for generating forecasts which has been discussed on the literature (see Wolfers and Zitzewitz 2004). Economists see them as an efficient aggregator of information by encouraging market participants to use various information sources and by exploiting the wisdom of the crowd effect. This effect was famously

described by (Galton 1907), who observed that average of all estimates of the ox weight at an exhibition was much more precise than any individual estimate. Markets essentially attempt to incorporate this effect. A well-known prediction market example is the Iowa Electronic Market , which has been used mostly to forecast elections in the U.S. But markets have been also used to forecast other political events as well, e.g. the outcome of referendums or even terrorism.[2]

For several reasons forecasting markets never really took off and still are of limited utility for policy making. One the hand this is due to the operating of prediction markets. One problem with market predictions is that they have shown to exhibit a favorite long shot bias. Small probabilities are overvalued and near certainty undervalued. This has been extensively discussed in the case of horse races, where people over proportionally bet on underdog horses (Thaler and Ziemba 1988). It was also shown to be a problem in financial markets (D. Bates 1991; Rubenstein 1994). Other operational problems of prediction markets are trading by desires (e.g. Forsythe, Rietz, and Ross 1999) and speculative bubbles. However, overall these problems tend to decrease with market size and sophistication.

More problematic for prediction market are its practical limitations: First, markets are not always feasible or even desirable. Take for instance the case of terrorism, where betting on these events create perverse incentives for conducting attacks. Or the case of asymmetric information, where strong insider knowledge exists and outsiders are basically discouraged to participate in the market. This is, for example, the case for many government decisions, where relevant forecasting information is only available to a small cycle of individuals. Second, prediction markets do often lack the necessary liquidity and number of market participants in order to produce meaningful predictions.[3]

A relatively new approach in forecasting international politics are forecasting competitions. In such a tournament participating individuals or teams are asked about the likelihood of future events. After the end of the forecasting horizon these forecasts are compared and evaluated. Forecasting competitions have some similarity to citizen election forecasts, where voters are asked which candidate or party they consider most likely to win their constituency (Murr 2011). Forecasting competitions combine the advantages of prior discussed approaches. Like intuitive predictions, they are hardly limited in their scope.[4]

[2]There was a DARPA research project on Policy Analysis Market (PAM), it was however stopped after public criticism.

[3]Betting markets might be a bit of an exception, but so far betting has not been used much as a political forecasting tool for events other than elections and referendums.

[4]The only serious restriction is, that the events under consideration are measureable in order to make subject to a forecasting competition.

They can incorporate reliable statistical models, if they are available. And like prediction markets, forecasting competition operate on the idea of aggregating the wisdom of the crowd and aggregating different information sources in order to generate forecasts.

In the field of international politics, this approach became prominent by the IARPA tournament, a geopolitical forecasting competition started in 2011 by the U.S. intelligence community. Different groups of academics were invited to participate in the project and compete against each other in providing forecasts about geopolitical events. In consecutive years, the winning team demonstrated that more accurate forecasts can be achieved by using two strategies: First, the extremization of individual judgements (Baron et al. 2014, V. Satopää and Ungar (2015)). Second, the exploitation of individual differences in forecasting skills and judgements (Atanasov et al. 2016, B. Mellers et al. (2015), V. Satopää, Pemantle, and Ungar (2017)). The first approach will not be discussed in detail here. Prior research on the second strategy has primarily looked at different predictors of forecasting success (B. Mellers et al. 2015; Poore et al. 2014). This paper is adding to this research by testing the replicability of prior findings and exploring new hypotheses.

The possibility of calibrating aggregated forecasts is a major advantage in comparison to prediction markets where market participants are usually anonymous and the individual leverage on the aggregated forecast is determined by the available capital of forecasters, which often comes from something else than their ability to forecast geopolitical events accurately. This is probably a main reasons why wisely aggregated forecasting competitions outperform markets (Atanasov et al. 2016).

Before I turn to the discussion of hypotheses, I want to touch upon the limitations of forecasting competitions. Limiting the understanding of successful forecasting to accuracy is their main weakness. One the one hand, forecasting is about more than just accurately describing the likelihood of events. It does also require exploring the unknown and identifying possible future events. In the forecasting competitions discussed here, this is not part of the tournament itself.[5] It ultimately requires recourse to other methods, e.g. expert opinion or strategic foresight (R. Popper 2009; Kosow et al. 2008; Bergheim 2009). On the other hand, accuracy might not be the most relevant quality of forecasts for policy decision making. For this other dimension like the impact or the possibility of early action might be more important.

There are, however, several reasons why accuracy should be considered of key importance for forecasting. First, the likelihood of future event is essential for decision making

[5]Here this is done by the tournament facilitator, which is discussed in more detail in Section 4.

from a normative point of view. Decision theory under uncertainty, and in particular its most used approaches (expected utility theory (Neumann and Morgenstern 1944) and subjective expected utility (Savage 1972)) presuppose the decision maker has an idea of the likelihood of the outcomes. Moreover, in practice decision makers tend to ask for a likelihood assessment when presented with possible future scenarios.[6] Second, improving accuracy is a process which has positive effects on the other aspects of forecasting (P. Tetlock and Gardner 2015). It creates good incentives for better forecasts in a broader sense by reducing ambiguity in forecasts, encouraging learning from mistakes and forcing the organizers of a forecasting competition to explore the "unknown". Third, more accuracy in foreign policy is possible and can be consistently utilized by analysts, e.g. by aggregating more information into single forecasts or discussing the implication of single pieces of information (Friedman et al. 2016).

3 Theory and Hypothesis

Forecasting competitions can be used to improve forecasts by identifying good forecasters, the environment in which they make good decisions and decision aids which can improve the quality of the judgement. Following this idea, the factors discussed in the paper can be divided along three categories: dispositional, environment-related and intervention. Dispositional factors are characteristics of the forecaster like abilities. They don't have to be fixed, but they should be at least stable on the short-term and measurable by a third party. Environment-related factors refer to the situation in which the decision is made. Intervention refers to treatments by outsiders, which aim at improving the quality of a decision.

Accepting the idea of individual differences in forecasting skills implies that there are individual dispositions which have direct implications on forecasting. In order to identify good forecasters it is therefore essential to know what characteristics can predict successful forecasting.

A good starting point for this is intelligence. It has been shown to be a good predictor for many other things like job performance (Ree and Earles 1992; Schmidt and Hunter 2004), socio-economic status (Strenze 2007), academic achievement (Furnham and Monsen 2009) and decision competence (Del Missier, Mäntylä, and Bruin 2012; Parker and Fischhoff

[6] Observation made by scenario and foresight experts which they expressed to me in a private conversation.

2005). Therefore, it can also be assumed to be a valuable predictor for forecasting success.

Intelligence can be defined in different ways. Generally, it is conceptualized either one-dimensional or multi-dimensional. In the one-dimensional approach a meaningful intelligence measure can be collapsed into one variable, while in the multi-dimensional approach intelligence is understood as a collection of different abilities.

B. Mellers et al. (2015) have argued that there are three relevant aspects of intelligence for geopolitical forecasting. They are (1) inductive reasoning (e.g. linking current problem and historical analogy), (2) cognitive control (e.g. override seemingly obvious but incorrect responses and engage in more prolonged and deeper thought) and (3) numeric reasoning (understanding mathematical dimension of a problem). In their research the different aspects were correlated to more accurate forecasts and they are correlated between each other. The first finding shows that intelligence measures can be used for predicting forecasting success and the second point supports the view that a one-dimensional intelligence concept is sufficient in this context. Similar results were derived by Poore et al. (2014), who found a strong correlation between various measures for analytical abilities and forecasting accuracy. Therefore it is reasonable assume these findings will be confirmed by the security policy forecasting tournament:

Hypothesis 1a: More intelligent individuals are more accurate forecasters

To test the hypothesis, it needs to be clarified how intelligence is measured. Various psychometric measures are available. B. Mellers et al. (2015) for example use three different methods: The Ravens Advanced Regressive Matrices method (Bors and Stokes 1998), the Cognitive Reflection Test (CRT) by (Frederick 2005) and a combined numeracy scale from Lipkus, Samsa, and Rimer (2001) and Peters et al. (2006). Poore et al. (2014) used self-reported SAT scores, sample GRE/SAT questions, subjective numeracy (Fagerlin et al. 2007) and, like B. Mellers et al. (2015), the CRT.

This paper uses a test which was not used for forecasting accuracy yet: The Berlin Numeracy Test (BNT) by Cokely et al. (2012). It is a relatively new psychometric scale that was in particular developed to access statistical numeracy and risk literacy. This makes the test in particular suitable for forecasting decisions involving probability judgements. Since it is especially suited to differentiate between individuals with higher education, it is also likely to perform better than the numeracy scale used by B. Mellers et al. (2015) which lacked differentiation. Another reason for the BNT is its length: The test consists only of four questions, making it suitable for one time forecasting tournament.

Intelligence is a common measure, but by far not the only dispositional factor of interest.

Other factors considered in this context include personality and cognitive styles (B. Mellers et al. 2015; Poore et al. 2014). It has been found that factors like openness to new experiences and active open-mindedness are good predictors for forecasting success. To keep the forecasting tournament within reasonable time limits for the participants, this research did not attempt to replicate these results. Instead, the paper scrutinizes a factor which was not tested before, but was raised by P. Tetlock and Gardner (2015, 226) in a side note: The interference of moral judgements with analytical judgments. In his book Tetlock claims superforecasters, unlike other forecasters, can separate analytical judgements from moral judgements. This is in particular important when forecasters make judgements about events for which they have a strong moral position. In such cases many people mingle the desirability of an outcome with the likelihood assessment. For example, if a person holds a strong moral opinion on the Syrian government and has a strong desire for it to be replaced by a more human-rights oriented government, Tetlock assumes this person to skew his or her probability judgment on the fall of the government towards the desired outcome.

This idea is similar to the social desirability bias where individuals over-report characteristics about themselves which they consider socially desirable (Dalton and Ortegren 2011) as well as to trading according to desires in prediction markets, where personal political preferences affect buying decisions (Forsythe, Rietz, and Ross 1999). Analogously, people are expected by Tetlock to overrate the likelihood of events they deem desirable.

But is this true? There are several ways this could be tested. First, the straight forward approach would be to ask respondents not only about the probability of an event but also about the desirability of it. The information could then be used to see whether it has any relationship to forecasting accuracy. But this approach does not help us further here. One the one hand the desirability of the events in question is unknown, although, this could theoretically have been asked in the forecasting tournament. One the other hand, for practical and normative reasons the accuracy of a forecaster should be predictable without knowing her or his moral opinion on the subject. Practically, it would be unfeasible as one would always have to ask forecasters about probability and desirability of events they forecast. This would increase the workload and induce fatigue. Normatively, it is undesirable to force forecasters to reveal their personal opinions as many of them are likely to work in highly politicized environments.

Another way would be to rank the questions according to their morality and see whether forecasters accuracy is related to the morality level of a question. However, I am not aware of any morality scale which could be applied to forecasting questions and which

is sufficiently universal. To the contrary, it is plausible to have people disagree about the morality of geopolitical events. For example, many in the West would like Assad to lose power of the Syrian government while a Syrian Alawite likely will have a very different view on this. Moreover, assigning morality levels to questions introduces a measurement problem: It cannot be distinguished whether the source of the inaccuracy stems from the uncertainty of an event in question or its morality.

Hence, we might have to rely on a less direct test of the relationship of moral and analytical judgement. The (psychological) theory of moral judgement might be a promising starting point for this. According to Kohlberg (1958) the moral development of humans can be ordered on a scale. This scale reflects how sophisticated moral justifications of judgements are. In this context, the concept of moral competency was developed. It describes how people can disentangle moral decisions along this scale and see consistently the differences between them. This has some similarity of that we are interested in the context of forecasting questions. If moral competent individuals are better able to differentiate between different moral justifications, we might also assume that they can better differentiate their moral and their analytical judgement in the context of a forecasting question. In order to test this, the following hypothesis is used:

Hypotheses 1b: More moral competent individuals are more accurate forecasters

More precisely, moral competency can be defined as "the ability of a subject to accept or reject arguments on a particular moral issue consistently in regard to their moral quality even though they oppose the subject's stance on that issue" (Lind 2008, 200). It can be contrasted with opinionated judgements, which are intuitive and emotional reactions to the content at hand and much of what Tetlock understands by moral assessment. Moral competency hence captures the idea that individuals are willing to consider counterviews despite their own, potentially strong, view on the issue.
In this regard moral competency has similarities to the active open-mindedness measure (Baron 2007), where respondents are asked whether they would consider other opinions and which has been shown to be a good predictor of forecasting success (B. Mellers et al. 2015). But it goes a step further: Instead of relying on a self-assessment, it actually tests whether different arguments are considered in the face of a moral issue. The hypothesis is therefore:

Moral competency can be measured with the moral competency test (MCT). The test confronts respondents with two moral dilemma situations. To each story the participants have to answer 12 questions on different justification for the described acts. The answers

are used to compute a competency score for each respondent. The score is not based on correct and wrong answers, but reflects a ratio between different parts of the answers. [7]

Individual differences in forecasting accuracy are also determined by differences in the decision environment. Unlike early decision theory assumed, empirical research has shown that context matters (e.g. D. Kahneman and Tversky 1974). The list of possible factors is long, ranging from the number of forecasters, the incentives to the opportunity of deliberative practice (Arkes 2001; Ericsson, Krampe, and Tesch-Römer 1993; D. Kahneman and Klein 2009). In this paper only the simplest environmental factor will be tested: Time used for answering the forecasting questions.

Time used for forecasting is ultimately a choice of the forecaster. It can, however, be influenced by explicitly making time slots available or freeing forecasters from other tasks. To justify such choices, it would be necessary to know the relationship between time and forecasting accuracy. There are good reasons to belief that more time will also lead to more accurate forecasts. First of all, time is necessary in order to go beyond intuitive thinking and to engage in analytical thinking about the question at hand. The two different ways of thinking are often described as 'system 1' and 'system 2', where system 1 stands for fast intuitive judgements and system 2 for slow reflective thinking (Evans and Stanovich 2013; D. Kahneman 2013). There is also a second reason why more time might lead to more accurate forecasts: Spending time on a question does allow gathering more information. Hence, the use of time should indicate whether a decision was informed or not.

However, there are two reasons why this might not be the case. First, the forecasters might use the time for other things. Haran, Ritov, and Mellers (2013), for example, have argued that acquiring new information depends on other characteristics like active open mindedness. But this is unlikely to be a problem in this forecasting tournament as participants would have no reason to spend more time on the questions and move on to the next section of the tournament. Second, the forecasters might have different levels of pre-knowledge. Hence, some participants might need more time to grasp the context of questions while

[7] Basically, a multivariate analysis of variance (MANOVA) is used here which measures how individuals disaggregate different moral stages. Based on the instructions of the test (Lind, 2008) I reengineered the underlying formula for the score:

$$c = \frac{\frac{1}{4}\sum_{i=1}^{6}(\sum_{j=1}^{4} x_{ij})^2 - (\frac{1}{24}\sum_{i=6}^{6}\sum_{j=1}^{4} x_{ij})^2}{\sum_{i=1}^{6}\sum_{j=1}^{4} x_{ij}^2 - (\frac{1}{24}\sum_{i=6}^{6}\sum_{j=1}^{4} x_{ij})^2} \cdot 100 \quad (1)$$

where $i \in 1, ..., 6$ stands for the stage and $j \in 1, ..., 4.$ for the section (pro-Worker, contra Worker, pro-Doctor, contra-Doctor). The score is between 0 and 100 and is sometimes categorized as follows: very low (1-9), low (10-19), medium (20-29), high (30-39), very high (40-49) and extraordinary high (above 50) (Lind, 2008, p. 200). For further information check the associated website.

others can rely on their extensive political pre-knowledge. As will be later described in more detail, this was counteracted by ensuring a wide span of questions. Moreover, going back to the first explanation for the link between time and accuracy: Even individuals with pre-knowledge will have to switch between the two mental modes, which should again be captured by the time spend on the questions.

But the relationship between time and accuracy is unlikely linear. From the view of a mental system shift, there is no theoretical reason why more time should increase accuracy once the shift of mental systems took place. From the informational point of view, over time the value of new information decreases as it will have less and less implications for the judgement and the costs of gathering will increase as it will be harder to find new additional information. For this reason, it is reasonable to assume that the marginal value of more time decreases:

Hypothesis 2: The marginal added value of time spend on forecasting is positive and decreases over time

In order to verify the hypothesis, the time participants used to answer the forecasting questions is measured. The validity of the measurement is ensured by treating other questions asked over the course of the forecasting completion in different sections and thereby excluded from the time measurement for this hypothesis. Moreover, the time used for forecasting is cross-checked with self-reported time use.

Finally, differences between individual forecasters can be the result of outside interventions. In forecasting this could be achieved by treatments improving analytical judgements (Soll, Milkman, and Payne 2015; Larrick 2004). Possible interventions include minor decision aids (Kretz 2015), feedback (Benson and Önkal 1992), exposure to multiple perspectives (Ariely et al. 2000; Herzog and Hertwig 2009), exposure to historical analogies (Lovallo, Clarke, and Camerer 2012), decomposition of problems into subsets (Fischhoff, Slovic, and Lichtenstein 1978), explicit consideration of contradictory evidence (Koriat, Lichtenstein, and Fischhoff 1980) and probabilistic training (B. Mellers et al. 2014). They can be tested in a forecasting tournament. B. Mellers et al. (2014), for example, tested the effect of scenario and probabilistic trainings and found them to have a long-term positive effect on forecasting accuracy.

However, for the forecasting tournament in this paper the focus will be on minimal interventions as the scale of the tournament is rather small. It is therefore more suited to test mild interventions aiming at debiasing. This could, for example, be achieved by providing forecasters with analytical tools. Kretz (2015) has done some research on mild

decision aid interventions and found that most analysts tend to disregard decision aids which require some effort, at least after some time. Kretz sees the reason for this in the additional mental capacities needed to apply the decision aids, which distracts the analysts from the actual problem at hand. In his research only the mildest intervention proved to have a significant effect on the judgement quality.

Following this idea, I decided to use a decision guide as treatment. The guide is inspired by Tetlock's description of how "superforecasters" approach forecasting question (B. Mellers et al. 2014). Theoretically speaking, it advises the forecaster to find a reference class for the forecasting question and then use Bayesian updating to adjust the base probability with other information.[8] Practically, this means forecasters were advised to identify a base rate for event in the forecasting question (outside view) and add or subtract incrementally probability points from this depending on the nature of the available information (inside view). Basically, this is a decision heuristic. The decision guide might improve forecasting accuracy for two reasons: First, the decision heuristic is based on standard theory for decision under uncertainty. This theory reflects the view most decision theorists have on how such questions should be addressed. Second, the decision heuristic of out- and inside view has been empirically successful for other types of forecasting decisions, e.g. for company revenues (Lovallo, Clarke, and Camerer 2012).

Hypothesis 3: A decision guide increases the accuracy of forecasting

In order to test this, all participants are assigned randomly, in about equal shares, to a treatment or control group. The treatment group is provided with a short (ca. 150 words) decision guide which describes the idea of outside and insight view in simple words and illustrates it with an example.[9] At the end of the guide, they were asked to fill out a small check box on whether they have read the guide.

4 The Forecasting Tournament

The forecasting tournament took place from 06.-12. February 2017 and forecasters were asked to consider possible events happening between 12th of February and the 24th of April 2017. The forecasters came either from the Master of International Relations program

[8] The reference class raises fundament issues about interpreting probabilities, as for single events the classical frequency view of probability does not work (e.g. K. Popper 1959). Hence, the forecasters have to pick a reference class based on some similarity criteria.

[9] The full text of the guide is available in the online appendix

at Hertie School of Governance Berlin or were recruited externally via mailing lists of relevant study programs, associations working in the field of international relations or by word of mouth. The students had to do the tournament as a homework while the others participated voluntary. The group was further complemented by forecasters from Amazon Mechanical Turk, who were paid for their participation. In total, 214 forecasters provided valid answers, they had an average age of 31.4 and 45.3 percent of them were female.

The forecasting tournament was conducted with a one off online survey. The survey consisted of three parts: First, the forecasters were asked question batteries of psychometric measures on intelligence and moral judgement. In the second part, the participants answered 24 forecasting questions on various security policy related events.[10] In each forecasting question the participants were asked to provide a judgement of how likely they thought the event is, expressed in probability. Finally, in the third part, forecasters reflected upon their forecasting and provided some demographic information about them.[11]

The forecasters were informed that their background information will be handled confidentially and in case of the university group that their forecasting judgments will be accessible to their fellow students. The second notice had the intention to incentive the students to give serious consideration to their answers by creating a competitive environment. In case of the voluntary participants this was less important as their participation already indicated intrinsic motivation. For the Mechanical Turk users the survey used attention checks, which they had to pass in order to receive the payout. All participants were also explicitly informed to use all information sources the deem relevant and spend as much time on the questions as they need. The participants were not informed whether to work individually or in teams as this lies outside of the control of the research design, but 9.8 percent said they answered the questions not alone. In total 231 individuals participated in the security policy forecasting tournament. However, only 214 responses are used as some participants had signs of not taking the survey serious (failing attention checks, unrealistically short time used for participation) and or submitted their forecasts after February 12th, 2017.

The questions were all related to security policy and selected in a multi-stage procedure. In the first step, I selected conflict regions which might be subject to changes in the short time horizon under considerations. Then I drafted questions by surveying reports from international organizations, think tanks, governments, NGOs and media outlets on recent developments in the conflicts. Among these sources were reports by the International Crisis

[10] A full list is the questions is available in the online appendix

[11] The survey was implemented with Qualtrics.

Group, the German Institute for International and Security Affairs (SWP Berlin), German Institute of Global and Area Studies (GIGA), Brookings Institute and the Carnegie Center. Finally, the draft questions were sent to a few researchers and forecasting experts for feedback and their recommendations were integrated in the final forecasting tournament.

The questions were all binary and the possible answers "yes" or "no". Most questions covered possible events in the whole time period between February 12th and April 24th, 2017. For each question the forecasters had to specify a probability to indicate how likely they expected the event to be. One question was, for example, "Will IS claim responsibility for another attack with a truck inside the European Union by 24. April 2017?". The questions have to be precise and measureable. How difficult this is, one can see by the mentioned question. On March 22th, 2017 the Westminster Attack happened in the UK. The attacker used a SUV to attack and kill several people in London. The question was intended to capture such events, but the term 'truck', literally understood, does not include SUVs. This is a fundamental problem about outlining events which did not happen yet: There will be aspects which were not anticipated correctly. In this case, the type of the car. How to respond in such cases? Here the event was nevertheless seen as a 'yes' reply to the question. First, the question was intended to capture such events and the use of an SUV instead of a truck does not make it fundamentally different. Second, suppose one would ask the participants whether their expectation explicitly excluded the case of SUV being used the likely answer would be no. However, there is no fixed rule for these borderline cases and they need to be decided on case to case basis.

Selecting the questions illustrates the distinction between two challenges in forecasting: Sampling and accuracy. In this research design sampling is done by the organizer of the forecasting competition while participants are solely dealing with the issue of accuracy. Ideally, one would also include sampling into a competition and testable format, but samples of different possible future events are hard to compare and therefore they cannot easily be made part of a competition.

As a good sample of possible geopolitical events is crucial for the forecasting tournament, the forecasting questions had to satisfy a number of criteria. First, the questions should neither concern events which have almost no chance of happening nor almost certain events. It is difficult to select a highly unlikely event as they are numerous and can have all kinds of realizations. An example for such an event is the start of the Arab Uprising in 2011 after Mohamed Bouzid set himself on fire. Neither should almost certain events be subject to a forecasting competition. An example for such a question would be whether the German federal elections will take place in September 2017. To some degree such a

question is a just the flipside of the highly unlikely event, but without specifying what this interruptive event could be. But again, choosing a relevant almost certain event for a forecasting competition will become an arbitrary choice. Moreover, remote and almost certain events will cause clustering of forecasts along the extreme values by participants in the forecasting competition. This would make it harder to distinguish successful forecasters from unsuccessful ones.

Second, the questions should cover various regions. The results might be biased if individual participants have special knowledge about a region which is overly represented in the competition. To further reduce the effects of narrow expertise, similar questions were also avoided. Ideally, the questions should also cover a wide range of policy fields. However, as the security policy forecasting tournament was conducted in collaboration with a university course at Hertie School of Governance, the topics were restricted to content of this course. This should not be a problem, as the prior mentioned considerations already introduce diversity into the questions and even within the field of security policy there is a wide range of possible topics.

Third, events for the forecasting competition should be relevant for policy makers and a large group of people. Relevance implies that the event has an impact on policy makers. The impact dimension excludes forecasting questions like the music played at the inauguration of a head of state. Relevance in this research was ensured by selecting events or indicators which would be discussed or considered by international organizations, governments and policy-oriented research institutions.

Even though these criteria guided the selection of the questions, a few limitations had to be taken into account. First of all, language restricted the range of possible events. Only events which would be reported in English language were selected for the competition. It reduces the problem of forecasters benefiting from the knowledge of certain languages. This could, for example, be the case with Spanish as some events in Latin America most information would be in Spanish. However, the more significant events are the more likely there is also sufficient information in English available. Second, the events were chosen on the basis of possibly getting international media attention. On the one side, this reduces the barrier for participants as it limited the scope of the questions to topics they might at least generally familiar with. On the other side, it keeps the workload for tracking questions reasonable. However, this does also exclude many possible questions. For example, funding decisions in international organizations are of policy relevance and might have severe implications, but information on them are hardly available. Third, for the events should be reliable information available. In the field of security policy this can

be difficult as information are inherently subject to the conflict dynamics and in many conflict areas almost any reliable information is hard to get by. Take for instance the conflict in the Democratic Republic of Congo or even Syria, where smaller incidences are rarely reported, and even if, cannot independently be confirmed.

In order to assess the quality of forecasts, they have to be scored. A common method is based on the Brier score, which was originally proposed in the context of weather forecasting (Brier 1950). Generally speaking, the Brier score indicates the distance of the forecast to the truth. More precisely, the Brier score is the squared error of a probabilistic forecast. To calculate it, the forecast are expressed on the range between 0 (0%) and 1 (100%). The realized events are coded either 0 (if the event did not happen) or 1 (if the event did happen). For each answer option, the difference between the forecast and the correct answer is squared and added. It can be expressed with:

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{R}(p_{ik}-o_{ik})^2 \tag{2}$$

N stands for the number of events, R is the number of possible classes the event can fall, p is the probability forecast and o the realized outcome. The Brier score can evaluate questions with more than two possible outcomes ($R > 2$), but in this paper only binary events are considered ($R = 2$).[12] The best (lowest) possible Brier score is 0, and the worst (highest) possible Brier score is 2. The Brier score is a proper scoring function which means that participant cannot improve their score by reporting a different probability from their actual belief. The participants' Brier score are computed by averaging the Brier scores across the questions.[13]

5 Results

The research design aims at identifying different factors of individual forecasting success. For this, measurable differences between the individuals are a prerequisite. In the case of a forecasting tournament, this can be verified by looking at the distribution of Brier scores

[12] As the competition only includes binary events, a simpler version of the Brier score (which is equivalent to the squared error) would also be sufficient. But to make comparison to the Good Judgement easier the multinomial version of the Brier score is used here. The simple Brier score can easily be computed by dividing the multinomial Brier score by two.

[13] In contrast to B. Mellers et al. (2015) the scores don't have to be normalized as the participants had to answer all questions and could not self-select the questions they thought to be easiest.

(Figure 1). Since the distribution ranges from 0.31 to 1.03 with a mean score of 0.56 we have enough variance for further testing.

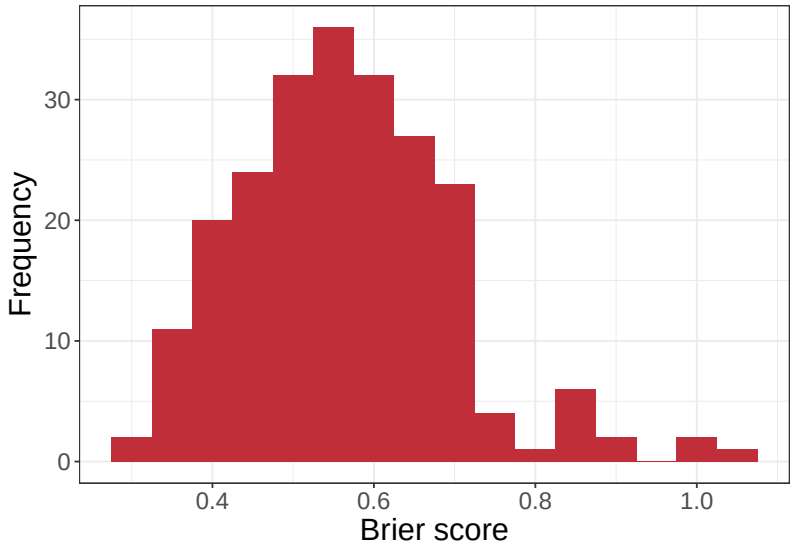

Figure 1: Brier score distribution

Before turning to the hypotheses, it makes sense to see how the participants performed in comparison to a simple statistical benchmark. This gives us a picture of the overall forecasting ability of the participants. A standard benchmark is the comparison of the realized forecasting scores to a situation with uniformly random distributed outcomes (B. Mellers et al. 2015). The average Brier score with random events is 0.65, while the actual average Brier score of the participants was 0.56 (t(213) = -10.49, p < 0.001).[14] Hence, the forecasting crowd performed significantly better than the benchmark. This supports the view of forecasting optimists: To some degree forecasting seems possible.

A second and more intuitive measurement of the overall forecasting accuracy of the participants is the proportion of questions where forecasters with their forecasts were on the correct side of 50% (B. Mellers et al. 2015, 6). The measure counts the forecasts above 50% for event which happened and forecasts below 50% for events which did not happen and divides them by the total number of forecasts. The perfect score would be 100%, a score corresponding to chance 50%. The forecasting average of the tournament participants was 56%. Again, we can use the t-test to measure whether the difference to

[14] In this context a one-sided t-test is used to see whether the forecasters performed better: $H_0 : \bar{b} = b_rand$ and $H_A : \bar{b} < b_{rand}$. The Brier score for random events was computed by computing the expected Brier score for each question / individual and taking the average: $\frac{1}{24 \cdot 214} \sum_{q=1}^{24} \sum_{i=1}^{214} (p_{qi}^2 + (1 - p_{qi})^2)$ with probability forecast $p \in [0,1]$, individuals $i \in \{1, ..., 214\}$ and questions $q \in \{1, ..., 24\}$.

Statistic	N	Mean	St. Dev.	Min	Max
Brier score	214	0.56	0.13	0.31	1.03
BNT Score	214	2.06	1.39	0	4
MCT Score	185	0.27	0.19	0.00	0.82
Forecasting time in min	212	12.22	19.96	0.71	184.91
Total time in min	214	201.03	707.92	3.17	4,640.80
Age	214	31.44	9.17	19	68

Table 1: Descriptive Statistics

the chance score is significant ($t(213) = 7.27$, $p < 0.001$).

Like for the first benchmark, this measurement indicates that forecasters performed better than random guessing. However, it does also illustrate that on average the forecasters are just slightly better than chance. For comparison: In the Good Judgement Project (B. Mellers et al. 2015, 6) forecasting competition the share was 75%, indicating that their forecasters crowd performed better.[15]

Having established the performance of the crowd, the focus can now turn to factors behind the forecasting success of individual forecasters. Starting with the first dispositional factor: intelligence. The forecasters mean score at the Berlin Numeracy Test (BNT) score was 2.06 (at a range from 0 to 4), which is slightly above the 1.6 average score Cokely et al. (2012) found for Berlin university students.[16] As the distribution of BNT scores illustrates (Figure 2), the test is able to discriminate between the participants. It assigned the forecasters to five score levels, each of which is roughly equal in size.

[15] A third measure for the overall performance of the crowd would be to compare the average score of the crowd to the performance of 50% guess for each question, which would basically assume the decision maker is ignorant to any information. A 50% guess for each question is equivalent to a Brier score of 0.5. Hence, according to this measurement the crowd actually performed worse than a simple 50% guessing strategy and the crowd's performance looks less favorable compared to the other measures. But this is not a problem, as the primary focus here is to understand the individual differences between forecasters and why some managed to outperform the rest.

[16] The forecasting tournament used the four item 'paper and pencil' version of the test. For this version there is no general population score for comparison available. However, in principle it would be possible by adjusting the score to the to the more commonly used adaptive test format. For further details see Cokely et al. (2012).

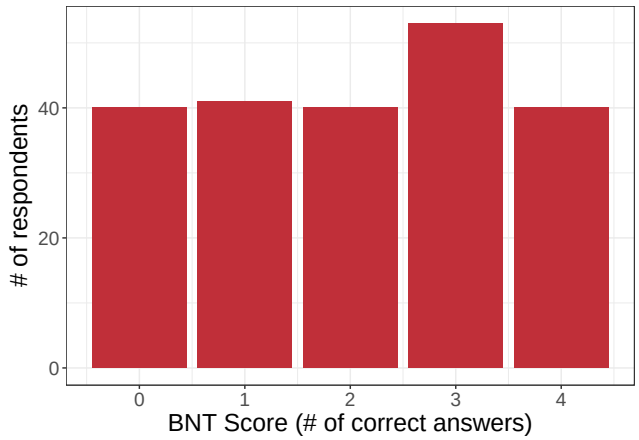

Figure 2: Berlin Numeracy Test score distribution

In order to understand the relationship between intelligence and forecasting accuracy, the Pearson correlation coefficient is informative. It computes the direction of the relation, indicated by the sign (+ or -), and the magnitude of the effect on a range from 0 to 1. The correlation between the BNT score and the Brier score is: r = -0.19, t(212) = -2.8, p = 0.006. With the t-test it can be further assessed whether the result is significantly different from no relationship between both variables ($r = 0$). [17] As expected the correlation is negative, which implies that more intelligent individuals tended to be more accurate forecasters (expressed in a lower Brier score). This supports the first hypothesis (1a) in line with prior findings in the literature. However, the relationship is not very strong. There can be several reasons for this: First, the relation between intelligence and forecasting accuracy might be less strong than previously argued. This is, however, rather unlikely, as intelligence was a strong predictor of forecasting success in more sophisticated research designs (B. Mellers et al. 2015; Poore et al. 2014). Second, the BNT test could be not valid and therefore not measure what it claims to. Again, this is rather unlikely as it has been extensively tested in various settings (Cokely et al. 2012). Nevertheless, one might argue that it measures an intelligence dimension which is less relevant for forecasting than other intelligence aspects. However, BNT score are correlated with other intelligence measures (Cokely et al. 2012) and there is no plausible reason why risk literacy should not matter for forecasting while other intelligence measures do. Third and most likely: The forecasting success in the tournament reflects a mix of skills and luck, which is skewed towards luck. This was already indicated by the rather moderate performance of the crowd against standard

[17] Here the following t-test statistic is used: $t = r\sqrt{\frac{n-2}{1-r^2}}$. Note that the Pearson correlation test treats the BNT score as a continuous variable, implying that the intelligence difference between the score levels is about equal.

benchmarks. The reason is probably the one-off nature of the forecasting tournament. In contrast to the forecasting projects of B. Mellers et al. (2015) and Poore et al. (2014) participants have little chance to incorporate feedback and improve their forecasting skills under these conditions. If this holds true, it is likely to be reflected as well in the remaining hypotheses testing.

	Brier score	BNT score	MCT score
Brier score			
BNT score	-0.19**		
MCT score	-0.17*	0.28***	
log(time)	-0.21**	0.28***	0.29***

Table 2: Correlation Table

Having this in mind, I can now turn to the second dispositional factor under scrutiny: moral competency. The Moral Competency Test (MCT) resulted in scores ranging from 0 to 0.82. Similar ranges were also found in other groups (Lind 2008). The moral competency score is not available for all participants since the completion of the moral competency test questions were not obligatory. However, the missing values are largely due to forgetting to answer a sub-question than to categorical non-replies.[18] Hence, the scores should not have systematic non-response bias. In order to assess the hypothesis, the Pearson coefficient is used to evaluate the relationship between forecasting accuracy and moral competency: r = -0.17, t(183) = -2.27, p = 0.025.

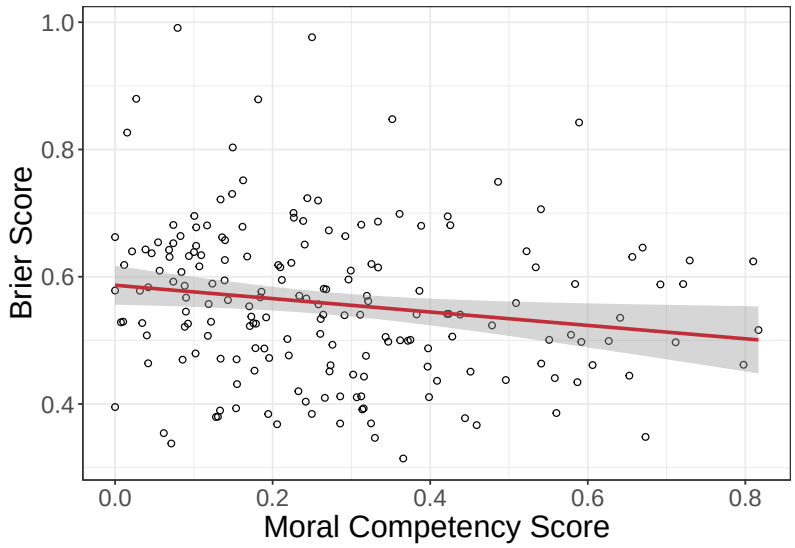

Figure 3: Scatterplot Moral Competency and Brier Score

[18]There is no procedure for calculating the MCT score with missing answers.

The correlation between moral competency and forecasting accuracy indicates that individuals with a higher moral competency score forecasted slightly more accurately. However, the effect is weak and not significant. Hence, it is not possible to draw a solid conclusion on the relationship between moral competency and forecasting accuracy. Like for the intelligence hypothesis, the role of luck versus skill might skew the result. However, in the case of moral competency one other major reason might also play a role: The validity of the MCT test for the hypothesis. The test is a valid measure to see whether people consistently evaluate moral questions on the basis of their moral quality when faced with difficult decisions (Lind 2008, 200). In other words: It looks whether individuals are considering different moral reasons for actions despite having a moral position for themselves. In the context of the forecasting tournament the questions whether the moral opinion of the forecasters affects their analytic judgements. This is not exactly the same. Nevertheless, there is a similarity in both tasks: In both cases the decision makers have to consider contrary views and incorporate them into their judgement. Hence, they have to be willing and able to incorporate new information while having a personal stand on the issue. The link between moral and analytical judgements can, however, only be disentangled with further research.

In hypothesis 2 the role of decision time for forecasting accuracy was tested. The median participants spend 24.6 minutes for participating in the survey and 6.4 minutes on answering the forecasting questions. Hence, a large share of the participants did not spend much time on the individual questions but relied on their intuitive judgement.[19] But this is not true for all participants, as some spend a considerable amount of time on forecasting. Hence, there is sufficient variance among the participants for the hypothesis. Since hypothesis 2 assumes the marginal benefit of time to decrease, the logarithmized time for answering the forecasting question is used instead of the actual time. The Pearson correlation between the logarithmized time and forecasting accuracy is: $r = -0.21$, $t(210) = -3.16$, $p = 0.002$.[20]

The negative correlation coefficient indicates that more time spend on answering the question is related to more forecasting accuracy. Moreover, when we compare the correlation of the linear time and the logarithmized time, the later has a higher correlation coefficient and clearly significant. ($|-0.21| > |r = -0.1|$). Hence, this supports the view of a decreasing marginal return on forecasting as it was expected in hypothesis 2. To conclude: Forecasters who used more time performed better, but the added accuracy decreased as more time

[19] Overall, 66.4% of the participants said they used only or mostly intuition to answer the questions.

[20] Two extreme outliers were excluded as the recorded time was implausibly high, likely as a result of interrupting the survey in order to do something else.

they spend on the forecasting questions.

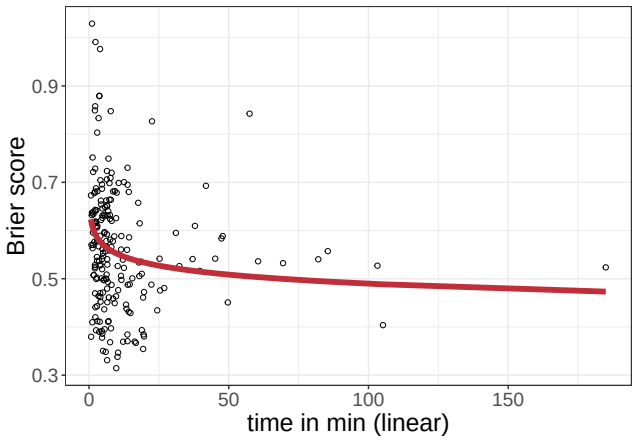

Figure 4: Forecasting time - Brier score scatterplot

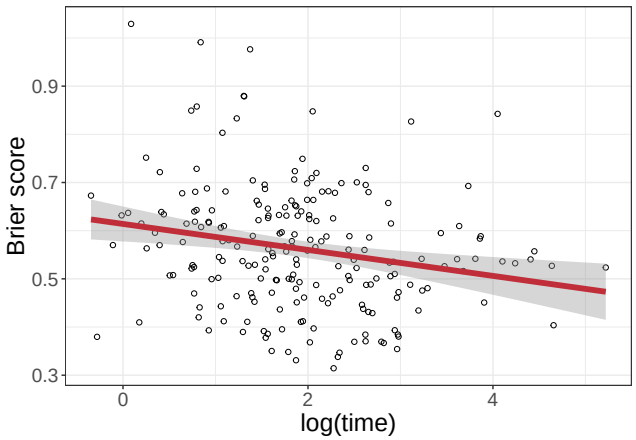

Figure 5: Log (forecasting time) - Brier score scatterplot

Finally, hypothesis 3 was about the effect of an decision aid intervention. The participants were randomly assigned to a treatment (n = 101) or control group (n = 113). The treatment group was presented with a decision guide, while the control group was not. To test whether the analytic guide had any impact on the forecasting accuracy the mean Brier scores of both groups is used. The mean Brier score of the treatment group is 0.56 and of the control group 0.57.

There is hardly any difference between the two groups ($t(212) = -0.73$, $p < 0.2317$). Hence, it is unlikely that the intervention had any impact on forecasting accuracy and there is no support for the intervention hypothesis (3). There are several possible reasons for the failure of the decision guide:

First, participants might have ignored the decision guide. This might in particular be true for forecasters who only spend a few minutes on answering the questions. Moreover, applying a decision methodology requires cognitive effort from decision makers (Kretz 2015, 68ff) and disrupts the train of thoughts (Hernandez and Preston 2013). When faced with difficult analytical tasks, like forecasting questions, decision makers might have used their mental capacities rather for information processing than for their methodological approach. The intervention tried to account for this with a minimal preventative measure. The forecasters had to indicate with a check box whether they have read the guide. Check boxes have been shown to be an effective nudge for analysts to increase their attention and are commonly used, e.g. by airline pilots or marine crews, see Kretz (2015, 33ff.). All participants in the treatment group indicated that they have read the guide. So what happened? To see whether the treatment lead to any behavioral change, we can check whether it had any effect on the time used for answering the forecasting questions. And this is the case, as the treatment group used 13.9 min and the control group 10.7 min. But these means might be the result of few extreme outliers. To compare whether the difference is more than a random occurence, the logarithmized times can be used.[21] Applying a t-test shows that the treatment group used more time if we use a significance level of 10%, but not for 5% (2.02 > 1.79, t(210) = 1.65, p < 0.0503). Even though this is not a clear result, it does also not rule out a small behavioral change from the decision guide.

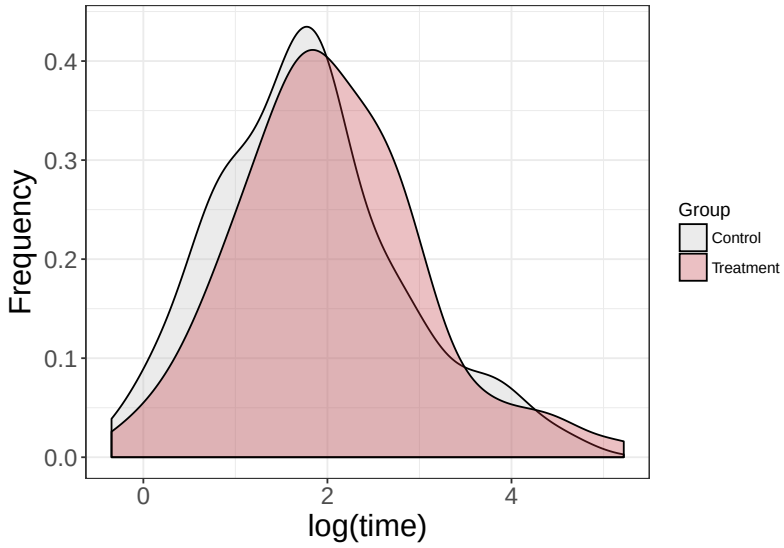

Figure 6: Distribution forecasting log(time) for treatment and control group

[21] Eye-balling the data shows that time in minutes used by the forecasters clearly does not follow a Gaussian distribution, but the logarithmized distribution resembles the classical Gaussian curve.

But even if the forecasters used the decision methodology, the lack of higher accuracy might also come from irrelevancy of the proposed methodology. However, this would contradict the research results by P. Tetlock and Gardner (2015) and his so-called superforecasters, who performed rather well with the methodology of the decision guide in the Good Judgement Project. Nevertheless, the problem might be a result of inappropriate application of the decision heuristic. In other words: In order to have measurable results in terms of forecasting accuracy, the forecasters need sufficient training in the methodology and a simple decision guide does just not provided enough learning experience. This is the most plausible explanation for the intervention failure, as for example B. Mellers et al. (2014) showed how a one-hour probabilistic training, which included the inside outside view methodology, could improve forecasting accuracy on the long term.

To conclude, similar to the 'list the hypotheses' or 'map the evidence methodology' (Kretz 2015) a simple decision guide can be added to a list of decision aids with no to limited effects for analytical decisions. However, what exactly is the reason for the failure is less clear.

6 Forecast aggregation

So far the focus of the paper was on the quality of individual forecasting decisions. However, having a group of forecasters enables us to generate forecasts by using the wisdom of the crowd effect. The simplest approach would be, like Galton (1907) did in the case of predicting the weight of oxen, to use the average probability forecasts of the participants. Compared to the average Brier score among the participants (0.56) the score from averaging the forecasts of each question is 0.43. Obviously, the crowd forecast performed significantly better than average participant in the tournament ($t(213) = 15.08$, $p < 0.001$).

But if we take into account what we know about individual forecasting decisions, the crowd wisdom can be aggregated in a more effective way. On the one hand, we can use individual predictors of forecasting success for combining the forecasts. On the other hand, extremizing aggregated forecasts can correct biases in aggregrated probability forecasts (Wallsten et al. 1997, Zhang and Maloney (2012)).

The individual differences between forecasters can be exploited in different ways when aggregating forecasts. A simple approach is to combine only forecasts of individuals who are expected to perform well. The forecast of this subgroup will then reflect both, individual forecasting skills and the wisdom of the crowd effect. The selection can be based

on the past performance of forecasters, as for example done by B. Mellers et al. (2014). Since in the security policy forecasting tournament the past performance of forecasters is unknown, this is not an option here. However, we can use the predictors of individual forecasting success to select a subgroup. As discussed in the previous section, these are the BNT score and the time used for forecasting. The choice of the direct cut-off point is somehow arbitrary. Here the subgroup is composed of the individuals who are the in the better half of the BNT score range and who spend sufficient time with forecasting.[22] Taking the average forecast of this of this selected subgroup results in a Brier score of 0.39, which is clearly better than the 0.43 score of the whole group. Theoretically, we could improve the result further by finding an ideal cutoff point for the subgroup. But then we risk overfitting the data and for illustration it is sufficient to show that the Brier score improves even with an arbitrary cut-off point.

Another approach is to weight the forecasts with the drivers of forecasting success. This approach has the advantage to avoid a cutoff point. The weights can be constructed in various ways, but it seems plausibe to assume that intelligence and the used time interact. In other words: A more intelligent forecaster should also be able to utilize the time more effectively. Therefore the weights are constructed by multiplying the BNT score and the logaritmized time ($w_i = bnt_i \cdot log(time_i)$).[23] Like for the subgroup aggregation, the weighted forecast performs better (0.41) than the unweighted average (0.43), but the improvement is rather small.

Apart from exploiting the individual differences, extremizing does also improve the aggregated forecasts. There are several reasons for this (Baron et al. 2014, V. A. Satopää et al. (2014)): First, the random errors are compressed at the ends of the 0 to 1 probability scale which pushes the average forecast towards 0.5. Second, individual forecasters draw from different sources. The diversity of information sources allows us to be more confident than the simple average forecasts suggests since the aggregate has a broader information base. This would not be the case, if all forecasters use the same information for their judgement.[24] Third, forecasts are underconfident because they only utilize particial information. Theoretically speaking, a forecaster would start from a 0.5 prior and

[22] More precisely: individuals which at least hat a BNT score of 3 and spend more time than the median forecaster on the questions.

[23] The logaritmized time is used to reduce the impact of outliers. It also better describes the relation between forecasting accuracy and time (Hypothesis 2).

[24] A commonly discussed example to illustrate this is President Obama's decision on whether to start a special operation to kill Osama Bin-Laden in Abbottabad. If his advisors used the same information sources, he would be best advised to average them. If they are from different sources, extremizing is advisable (e.g. P. Tetlock and Gardner 2015).

incrementally include information on average pointing in the direction of the best-formed probability forecast. However, as most forecasters don't use the time to incorporate all information they stop on the way to the best-informed forecast.[25]

For the implementation of the extremization this work orients at V. A. Satopää et al. (2014), who use a simple logit model to combine the forecasts. The logit transformation is convenient for extremizing as it maps the probabilties to a continuous domain where a 50% probability is equivalent to a log-odds value of 0. The systematic bias, which we want to correct for with extremizing, can be described with a single variable $a \in [0, \infty]$. $a = 1$ describes the case when the individual forecasts are the best-informed forecasts, while $a > 1$ captures underconfidence of forecasters. In principle, this representation is also applicable for overconfidence ($a < 1$). But former research has shown that forecasting crowds as a whole tend be underconfident (Baron et al. 2014). Mathematically, the relation between the best informed forecast (p) and the individual forecasts (p_i) can be descibed as follows:

$$y_i = log\Big(\frac{p_i}{1-p_i}\Big) = log\Big(\frac{p}{1-p}\Big)^{\frac{1}{a}} + \epsilon_i \tag{3}$$

In order to illustrate the potential of extremizing, we have to compute the corrected probabilities p_E for each question. This can be done with:[26]

$$\hat{p}_E(a) = \frac{\Big[\prod_{i=1}^{N}(\frac{p_i}{1-p_i})^{\frac{1}{N}}\Big]^a}{1 + \Big[\prod_{i=1}^{N}(\frac{p_i}{1-p_i})^{\frac{1}{N}}\Big]^a} \tag{4}$$

To illustrate the potential of extremizing, we can look for the level of bias correction (a) which minimizes the Brier score. For the security policy forecasting tournament the bias is a* = 1.24. Hence, as expected the forecaster crowd on average is underconfident. Extremizing does improve the forecasting accuracy to 0.407 from 0.43. Hence, extremizing can lead to similar accuracy gains as exploiting the individual differences between forecasters.

[25] Whether is really rational to start from a 0.5 prior when having no information has recently attracted some criticism, see e.g. Gilboa, Postlewaite, and Schmeidler (2009).

[26] This is an MLE estimator. For a more in-depth explanation see V. A. Satopää et al. (2014).

7 Discussion

To sum up, this reasearch has found the forecasting crowd to perform slightly better than random quessing but not as accurate as in other comparable forecasting competitions. The forecasting tournament supports the idea that intelligence and forecasting time, if diminished returns are considered, are useful predictors for forecasting accuracy. The data provides indicative support that moral competency is related to more forecasting accuracy, but findings are to weak to make a final conclusion on them. Finally, there is no support that the mild decision guide used as intervention has any measurable effect on forecasting accuracy. These insights can be used to aggregate forecasts more accurately by exploiting the individual differences between forecasters and using the wisdom of crowd effect.

These results have, however, should be enjoyed with care. First, the size of the tournament was relatively small compared to other similar projects. Having more questions and forecasters would make the results less sensitive to outliers and reduces the impact of single interpretation of the forecasting questions. Whether or not an event happened a few time needed some argumentation and with more questions the implications would likely balance each other on average. Second, the research design only allows to make limited claims. First off all, the tournament was a one off event which did not allow individuals to learn from their performance

be seen in the context that they are derived from a relatively small forecasting competition. The hypothesized relations should also play out more in a forecasting tournament where participants have the chance to learn. - only a one-off event

- participant number relatively small compared to other forecasting competitions in the field
- intervention limited due to online format
- small number of questions -> Measurement errors in the questions still have strong impact on the results.

References

Almond, Gabriel A., and Stephen J. Genco. 1977. "Clouds, Clocks, and the Study of Politics." *World Politics* 29 (04): 489–522. doi:10.2307/2010037.

Ariely, Dan, Randall H. Bender, Christiane B. Dietz, Hongbin Gu, Thomas S. Wallsten, Wing Tung Au, David V. Budescu, and Gal Zauberman. 2000. "The Effects of Averaging Subjective Probability Estimates Between and Within Judges." *Journal of Experimental Psychology: Applied* 6 (2 Access to Document Link to publication in Scopus): 130–47.

Arkes, Hal R. 2001. "Overconfidence in Judgmental Forecasting." In *Principles of Forecasting*, edited by Scott Armstrong, 30:495–515. International Series in Operations Research & Management Science. Boston, MA: Springer US. doi:10.1007/978-0-306-47630-3{\textunderscore }22.

Atanasov, Pavel, Phillip Rescober, Eric Stone, Samuel A. Swift, Emile Servan-Schreiber, Philip Tetlock, Lyle Ungar, and Barbara Mellers. 2016. "Distilling the Wisdom of Crowds: Prediction Markets Vs. Prediction Polls." *Management Science*. doi:10.1287/mnsc.2015.2374.

Baron, Jonathan. 2007. *Thinking and Deciding*. Cambridge University Press. doi:10.1016/B978-044450556-9/50070-8.

Baron, Jonathan, Barbara A. Mellers, Philip E. Tetlock, Eric Stone, and Lyle H. Ungar. 2014. "Two Reasons to Make Aggregated Probability Forecasts More Extreme." *Decision Analysis* 11 (2): 133–45. doi:10.1287/deca.2014.0293.

Bates, David. 1991. "The Crash of 87: Was It Expected? The Evidence from Options Markets." *The Journal of Finance* 46 (3): 1009–44. doi:10.1111/j.1540-6261.1991.tb03775.x.

Benson, P.George, and Dilek Önkal. 1992. "The Effects of Feedback and Training on the Performance of Probability Forecasters." *International Journal of Forecasting* 8 (4): 559–73. doi:10.1016/0169-2070(92)90066-I.

Bergheim, Stefan. 2009. "Zukunftsforschung Für Staaten: Vorbereitungen in Der Gegenwart." Zentrum für gesellschaftlichen Fortschritt. http://www.fortschrittszentrum.de/dokumente/2009-09_Zukunftsforschung_fuer_Staaten.pdf.

Beyerchen, Alan. 1992. "Clausewitz, Nonlinearity, and the Unpredictability of War." *International Security* 17 (3): 59–90. http://www.jstor.org/stable/pdf/2539130.pdf.

Bors, Douglas A., and Tonya L. Stokes. 1998. "Raven's Advanced Progressive Matrices: Norms for First-Year University Students and the Development of a Short Form." *Educa-*

tional and Psychological Measurement 58 (3): 382–98. doi:10.1177/0013164498058003002.

Brier, Glenn. 1950. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review* 78 (1). http://dx.doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Bueno de Mesquita, Bruce. 2009. *Predictioneer's Game: Using the Logic of Brazen Self-Interest to See and Shape the Future*. 1st ed. New York: Random House.

Choucri, Nazli. 1974. "Forecasting in International Relations: Problems and Prospects." *International Interactions* 1: 63–86.

Cokely, Edward, Mirta Galesic, Eric Schulz, Saima Ghazal, and Rocio Garcia-Retamero. 2012. "Measuring Risk Literacy: The Berlin Numeracy Test." *Judgment and Decision Making* 7 (1): 25–47. doi:10.1002/9781118170229.ch1.

Dalton, Derek, and Marc Ortegren. 2011. "Gender Differences in Ethics Research: The Importance of Controlling for the Social Desirability Response Bias." *Journal of Business Ethics* 103 (1): 73–93. doi:10.1007/s10551-011-0843-8.

Dawes, R., D. Faust, and P. Meehl. 1989. "Clinical Versus Actuarial Judgment." *Science* 243 (4899): 1668–74. doi:10.1126/science.2648573.

Del Missier, Fabio, Timo Mäntylä, and Wändi Bruine Bruin. 2012. "Decision-Making Competence, Executive Functioning, and General Cognitive Abilities." *Journal of Behavioral Decision Making* 25 (4): 331–51. doi:10.1002/bdm.731.

Dhami, Mandeep, David Mandel, Barbara Mellers, and Philip Tetlock. 2015. "Improving Intelligence Analysis with Decision Science." *Perspectives on Psychological Science : A Journal of the Association for Psychological Science* 10 (6): 753–57. doi:10.1177/1745691615598511.

Ericsson, K. Anders, Ralf T. Krampe, and Clemens Tesch-Römer. 1993. "The Role of Deliberate Practice in the Acquisition of Expert Performance." *Psychological Review* 100 (3): 363–406. doi:10.1037/0033-295X.100.3.363.

Evans, Jonathan St B. T., and Keith E. Stanovich. 2013. "Dual-Process Theories of Higher Cognition: Advancing the Debate." *Perspectives on Psychological Science : A Journal of the Association for Psychological Science* 8 (3): 223–41. doi:10.1177/1745691612460685.

Fagerlin, Angela, Brian J. Zikmund-Fisher, Peter A. Ubel, Aleksandra Jankovic, Holly A. Derry, and Dylan M. Smith. 2007. "Measuring Numeracy Without a Math Test: Development of the Subjective Numeracy Scale." *Medical Decision Making : An International Journal*

of the Society for Medical Decision Making 27 (5): 672–80. doi:10.1177/0272989X07304449.

Fischhoff, Baruch, Paul Slovic, and Sarah Lichtenstein. 1978. "Fault Trees: Sensitivity of Estimated Failure Probabilities to Problem Representation." *Journal of Experimental Psychology: Human Perception and Performance* 4 (2): 330–44. https://www.gwern.net/docs/predictions/1978-fischhoff.pdf.

Forsythe, Robert, Thomas A. Rietz, and Thomas W. Ross. 1999. "Wishes, Expectations and Actions: A Survey on Price Formation in Election Stock Markets." *Journal of Economic Behavior & Organization* 39 (1): 83–110. doi:10.1016/S0167-2681(99)00027-X.

Frederick, Shane. 2005. "Cognitive Reflection and Decision Making." *Journal of Economic Perspectives* 19 (4): 25–42. doi:10.1257/089533005775196732.

Friedman, Jeffrey, Joshua Baker, Barbara Mellers, Philip Tetlock, and Richard Zeckhäuser. 2016. "The Value of Precision in Probability Assessment: Evidence from a Large-Scale Geopolitical Forecasting Tournament Analysis." http://sites.dartmouth.edu/friedman/files/2016/11/Value-of-Precision-Nov-2016.pdf.

Furnham, Adrian, and Jeremy Monsen. 2009. "Personality Traits and Intelligence Predict Academic School Grades." *Learning and Individual Differences* 19 (1): 28–33. doi:10.1016/j.lindif.2008.02.001.

Galton, Francis. 1907. "Vox Populi." *Nature* 75 (1949): 450–51. http://galton.org/essays/1900-1911/galton-1907-vox-populi.pdf.

Gilboa, Itzhak, Andrew Postlewaite, and David Schmeidler. 2009. "Is It Always Rational to Satisfy Savage's Axioms?" *Economics and Philosophy* 25 (03): 285. doi:10.1017/S0266267109990241.

Goldstone, Jack A., Robert H. Bates, David L. Epstein, Ted Robert Gurr, Michael B. Lustik, Monty G. Marshall, Jay Ulfelder, and Mark Woodward. 2010. "A Global Model for Forecasting Political Instability." *American Journal of Political Science* 54 (1): 190–208. doi:10.1111/j.1540-5907.2009.00426.x.

Green, Kesten C., and J. Scott Armstrong. 2007. "Global Warming: Forecasts by Scientists Versus Scientific Forecasts." *Energy & Environment* 18 (7): 997–1021. doi:10.1260/095830507782616887.

Haran, Uriel, Ilana Ritov, and Barbara Mellers. 2013. "The Role of Actively Open-Minded Thinking in Information Acquisition, Accuracy, and Calibration." *Judgment and Decision*

Making 8 (3): 188–201. doi:10.1002/9781118403259.ch1.

Hernandez, Ivan, and Jesse Lee Preston. 2013. "Disfluency Disrupts the Confirmation Bias." *Journal of Experimental Social Psychology* 49 (1): 178–82. doi:10.1016/j.jesp.2012.08.010.

Herzog, Stefan M., and Ralph Hertwig. 2009. "The Wisdom of Many in One Mind: Improving Individual Judgments with Dialectical Bootstrapping." *Psychological Science* 20 (2): 231–37. doi:10.1111/j.1467-9280.2009.02271.x.

Kahneman, Daniel. 2013. *Thinking, Fast and Slow*. Farrar Straus and Giroux.

Kahneman, Daniel, and Gary Klein. 2009. "Conditions for Intuitive Expertise: A Failure to Disagree." *The American Psychologist* 64 (6): 515–26. doi:10.1037/a0016755.

Kahneman, Daniel, and Amos Tversky. 1974. "Judgment Under Uncertainty: Heuristics and Biases." *Science* 185 (4157): 1124–31. http://links.jstor.org/sici?sici=0036-8075%2819740927%293%3A185%3A4157%3C1124%3AJUUHAB%3E2.0.CO%3B2-M.

Kohlberg, Lawrence. 1958. "The Development of Modes of Thinking and Choices in Years 10 to 16." Ph.D. dissertation, University of Chicago.

Koriat, Asher, Sarah Lichtenstein, and Baruch Fischhoff. 1980. "Reasons for Confidence." *Journal of Experimental Psychology: Human Learning & Memory* 6 (2): 107–18. doi:10.1037/0278-7393.6.2.107.

Kosow, Hannah, Lorenz Erdmann, Robert Gaßner, and Beate-Josephine Luber. 2008. *Methoden Der Zukunfts- Und Szenarioanalyse: Überblick, Bewertung Und Auswahlkriterien*. Vol. 103. WerkstattBericht / Izt, Institut Für Zukunftsstudien Und Technologiebewertung. Berlin: IZT. https://www.izt.de/fileadmin/publikationen/IZT_WB103.pdf.

Kretz, Donald. 2015. "Strategies to Reduce Cognitive Bias in Intelligence Analysis: Can Mild Interventions Improve Analytic Judgment?" PhD Thesis, The University of Texas at Dallas.

Larrick, Richard P. 2004. "Debiasing." In *Blackwell Handbook of Judgment and Decision Making*, edited by Derek J. Koehler and Nigel Harvey, 316–38. Malden, MA, USA: Blackwell Publishing Ltd. doi:10.1002/9780470752937.ch16.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. "Big Data. the Parable of Google Flu: Traps in Big Data Analysis." *Science* 343 (6176): 1203–5. doi:10.1126/science.1248506.

Lewis-Beck, Michael. 2005. "Election Forecasting: Principles and Practice." *British Journal*

of Politics and International Relations 7: 145–64. 10.1111/j.1467-856X.2005.00178.x.

Lind, Georg. 2008. "The Meaning and Measurement of Moral Judgment Competence: A Dual-Aspect Model." In *Contemporary Philosophical and Psychological Perspectives on Moral Development and Education*, edited by Daniel Fasko and Wayne Willis, 185–220. Critical Education and Ethics. Cresskill, N.J.: Hampton Press.

Lipkus, I. M., G. Samsa, and B. K. Rimer. 2001. "General Performance on a Numeracy Scale Among Highly Educated Samples." *Medical Decision Making : An International Journal of the Society for Medical Decision Making* 21 (1): 37–44. doi:10.1177/0272989X0102100105.

Lovallo, Dan, Carmina Clarke, and Colin Camerer. 2012. "Robust Analogizing and the Outside View: Two Empirical Tests of Case-Based Decision Making." *Strategic Management Journal* 33 (5): 496–512. doi:10.1002/smj.962.

Mellers, Barbara, Eric Stone, Pavel Atanasov, Nick Rohrbaugh, S. Emlen Metz, Lyle Ungar, Michael M. Bishop, Michael Horowitz, Ed Merkle, and Philip Tetlock. 2015. "The Psychology of Intelligence Analysis: Drivers of Prediction Accuracy in World Politics." *Journal of Experimental Psychology. Applied* 21 (1): 1–14. doi:10.1037/xap0000040.

Mellers, Barbara, Lyle Ungar, Jonathan Baron, Jaime Ramos, Burcu Gurcay, Katrina Fincher, Sydney E. Scott, et al. 2014. "Psychological Strategies for Winning a Geopolitical Forecasting Tournament." *Psychological Science* 25 (5): 1106–15. doi:10.1177/0956797614524255.

Murr, Andreas Erwin. 2011. "'Wisdom of Crowds'? A Decentralised Election Forecasting Model That Uses Citizens' Local Expectations." *Electoral Studies* 30 (4): 771–83. doi:10.1016/j.electstud.2011.07.005.

Neumann, John von, and Oskar Morgenstern. 1944. *Theory of Games and Economic Behaviour*. Princeton University Press.

Norpoth, Helmut, and Thomas Gschwend. 2010. "The Chancellor Model: Forecasting German Elections." *International Journal of Forecasting* 26 (1): 42–53. doi:10.1016/j.ijforecast.2009.02.008.

Parker, Andrew M., and Baruch Fischhoff. 2005. "Decision-Making Competence: External Validation Through an Individual-Differences Approach." *Journal of Behavioral Decision Making* 18 (1): 1–27. doi:10.1002/bdm.481.

Peters, Ellen, Daniel Vastfjall, Paul Slovic, C. K. Mertz, Ketti Mazzocco, and Stephan Dickert. 2006. "Numeracy and Decision Making." *Psychological Science* 17 (5): 407–13. doi:10.1111/j.1467-9280.2006.01720.x.

Poore, Joshua C., Clifton L. Forlines, Sarah M. Miller, John R. Regan, and John M. Irvine.

2014. "Personality, Cognitive Style, Motivation, and Aptitude Predict Systematic Trends in Analytic Forecasting Behavior." *Journal of Cognitive Engineering and Decision Making* 8 (4): 374–93. doi:10.1177/1555343414554702.

Popper, Karl. 1959. "The Propensity Interpretation of Probability." *The British Journal for the Philosophy of Science* 10 (37): 25–42. doi:10.1093/bjps/X.37.25.

Popper, Rafael. 2009. *Mapping Foresight: Revealing How Europe and Other World Regions Navigate into the Future*. Vol. 24041. European Research Area. Research Policy. Luxembourg: European Commission; EUR-OP. http://www.forschungsnetzwerk.at/downloadpub/2009_efmn_mappingForesight_EU.pdf.

Ree, Malcolm James, and James A. Earles. 1992. "Intelligence Is the Best Predictor of Job Performance." *Current Directions in Psychological Science* 1 (3): 86–89. doi:10.1111/1467-8721.ep10768746.

Reyna, Valerie F., Christina F. Chick, Jonathan C. Corbin, and Andrew N. Hsia. 2014. "Developmental Reversals in Risky Decision Making: Intelligence Agents Show Larger Decision Biases Than College Students." *Psychological Science* 25 (1): 76–84. doi:10.1177/0956797613497022.

Rubenstein, Mark. 1994. "Implied Binomial Trees." *The Journal of Finance* 49 (3): 771–818. doi:10.1111/j.1540-6261.1994.tb00079.x.

Satopää, Ville A., Jonathan Baron, Dean P. Foster, Barbara A. Mellers, Philip E. Tetlock, and Lyle H. Ungar. 2014. "Combining Multiple Probability Predictions Using a Simple Logit Model." *International Journal of Forecasting* 30 (2): 344–56. doi:10.1016/j.ijforecast.2013.09.009.

Satopää, Ville, and Lyle Ungar. 2015. "Combining and Extremizing Real-Valued Forecasts." https://arxiv.org/pdf/1506.06405.pdf.

Satopää, Ville, Robin Pemantle, and Lyle Ungar. 2017. "Combining Probability Forecasts and Understanding Probability Extremizing Through Information Diversity." *Journal of the American Statistical Association*, to appear. https://www.math.upenn.edu/~pemantle/papers/Preprints/aggregation.pdf.

Savage, Leonard J. 1972. *The Foundations of Statistics*. 2d rev. ed. Dover Publiations. http://www.loc.gov/catdir/description/dover032/79188245.html.

Schkade, D. A., and D. Kahneman. 1998. "Does Living in California Make People Happy? A Focusing Illusion in Judgments of Life Satisfaction." *Psychological Science* 9 (5): 340–46.

doi:10.1111/1467-9280.00066.

Schmidt, Frank L., and John Hunter. 2004. "General Mental Ability in the World of Work: Occupational Attainment and Job Performance." *Journal of Personality and Social Psychology* 86 (1): 162–73. doi:10.1037/0022-3514.86.1.162.

Shanteau, James. 1992. "Competence in Experts: The Role of Task Characteristics." *Organizational Behavior and Human Decision Processes* 53: 252–66. http://www.sciencedirect.com/science/article/pii/074959789290064E.

Soll, Jack B., Katherine L. Milkman, and John W. Payne. 2015. "A User's Guide to Debiasing." In *The Wiley Blackwell Handbook of Judgment and Decision Making*, 924–51. John Wiley & Sons, Ltd. doi:10.1002/9781118468333.ch33.

Strenze, Tarmo. 2007. "Intelligence and Socioeconomic Success: A Meta-Analytic Review of Longitudinal Research." *Intelligence* 35 (5): 401–26. doi:10.1016/j.intell.2006.09.004.

Swets, J. A., R. M. Dawes, and J. Monahan. 2000. "Psychological Science Can Improve Diagnostic Decisions." *Psychological Science in the Public Interest : A Journal of the American Psychological Society* 1 (1): 1–26. doi:10.1111/1529-1006.001.

Taleb, Nassim Nicholas. 2007. "Black Swans and the Domains of Statistics." *The American Statistician* 61 (3): 198–200. doi:10.1198/000313007X219996.

Tetlock, Philip. 2005. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press.

Tetlock, Philip, and Dan Gardner. 2015. *Superforecasting: The Art and Science of Prediction*. Crown.

Thaler, Richard H., and William T. Ziemba. 1988. "Anomalies: Parimutuel Betting Markets: Racetracks and Lotteries." *Journal of Economic Perspectives* 2 (2): 161–74. doi:10.1257/jep.2.2.161.

Wallsten, Thomas S., David V. Budescu, I. Erev, and Adele Diederich. 1997. "Evaluating and Combining Subjective Probability Estimates." *Journal of Behavioral Decision Making* 10 (3): 243–68. doi:10.1002/(SICI)1099-0771(199709)10:3{\textless}243::AID-BDM268{\textgreater}3.0.CO;2-M.

Ward, M. D., B. D. Greenhill, and K. M. Bakke. 2010. "The Perils of Policy by P-Value: Predicting Civil Conflicts." *Journal of Peace Research* 47 (4): 363–75.

doi:10.1177/0022343309356491.

Ward, Michael D. 2016. "Can We Predict Politics? Toward What End?: Table 1." *Journal of Global Security Studies* 1 (1): 80–91. doi:10.1093/jogss/ogv002.

Wilson, Timothy, and Daniel Gilbert. 2005. "Affective Forecasting: Knowing What to Want." *Current Directions in Psychological Science* 14 (3): 131–34.

Wolfers, Justin, and Eric Zitzewitz. 2004. "Prediction Markets." *Journal of Economic Perspectives* 18 (2): 107–26. doi:10.1257/0895330041371321.

Zhang, Hang, and Laurence T. Maloney. 2012. "Ubiquitous Log Odds: A Common Representation of Probability and Frequency Distortion in Perception, Action, and Cognition." *Frontiers in Neuroscience* 6: 1. doi:10.3389/fnins.2012.00001.