**Supplementary Material**

Contents

1. Additional description of the Good Judgment Project’s methodology
2. Formal description of method for calculating rounding errors
3. Full results for analysis of returns to precision across time horizons
4. Analysis of returns to precision across question types
5. Additional description of the Good Judgment Project’s training module
6. Descriptive statistics for GJP forecasters’ individual attributes

I. Further description of the Good Judgment Project’s Methodology[[1]](#footnote-1)

The Intelligence Advanced Research Projects Activity hosted four forecasting tournaments from 2011 to 2015. Five university-based research groups competed to recruit forecasters, elicit predictions, and aggregate those predictions as accurately as possible. Each tournament lasted for approximately nine months of the year. The Good Judgment Project (GJP) was one of the five teams in this competition. GJP recruited forecasters via professional societies, research centers, alumni associations, science blogs, and word of mouth. Participation required a bachelor’s degree or higher and completion of a battery of psychological and political tests that took an average of two hours. Individuals who met the requirements were largely U.S. citizens (76%) and males (83%); their average age was 36. Almost two thirds (64%) had some postgraduate training.

*Questions*

To forecast, participants logged onto a website, www.goodjudgment.com. Each year, they were given approximately 125 geopolitical questions. They could select their own questions and answer as many as they wished. However, if they wanted to be paid, they needed to make forecasts for at least thirty questions. They were encouraged to update their beliefs as often as they wished before the close of each question. Questions were launched almost every week and remained open for an average of 102 days (range: 2-418 days). An example question was, “Will Italy’s Silvio Berlusconi resign, lose re-election/confidence vote, or otherwise vacate office before 1 January 2012?” Participants predicted the chance the event would occur, on a scale from 0% (*certain it will not occur*) to 100% (*certain it will occur*).

*Accuracy scoring and incentives*

All forecasters were given the goal of minimizing their average Brier score. Forecasters who met the minimum participation requirements of making predictions for at least thirty questions received $150 at the end of Year 1 and $250 at the end of Years 2, 3, and 4, regardless of their accuracy. Year 2 returnees were given a $100 bonus. Forecasters also received status rewards for their performance via leader boards that displayed the Brier scores for the top 10% of forecasters in each experimental condition and their own score. Those who served in teams (regular teams and superforecaster teams) also saw team Brier scores (defined as medians rather than means, to encourage harmony).

*Experimental Designs*

The Good Judgment Project had different experimental designs each year. Forecasters were randomly assigned to conditions, and if they continued for additional years, they usually remained in the same condition unless they became superforecasters. Superforecasters were the top 2% of forecasters selected from all conditions at the end of each year. These individuals were assigned to elite teams of roughly 15 members each and worked with each other.

In Year 1, GJP randomly assigned 2,400 forecasters to one of 12 conditions based on a 3 x 4 factorial design of Training by Elicitation. Levels of training were No Training, Probability Training, and Scenario Training, and levels of elicitation were Control (independent forecasters), Crowd Beliefs (independent forecasters who saw the distribution of others’ forecasts but were unable to communicate), and Teams (forecasters who worked in groups of 15-20 and were asked to justify the basis of their forecasts to each other).

In Year 2, GJP had approximately 1,600 forecasters. Almost all returning participants remained in the same condition from Year 1. New participants were assigned to one of 6 conditions based on a 2 x 3 factorial design of Training by Elicitation. Levels of training were No Training and Probability Training. Levels of Elicitation were a Control group (independent forecasters), Teams who worked in groups of 12 and were asked to justify their forecasts to each other, and Prediction Markets. Assignments to conditions were done such that numbers of forecasters in each condition would be roughly equal. GJP also had a 7th condition with just Superforecasters from Year 1.

In Year 3, GJP had 10 different conditions and approximately 3,000 forecasters. Four conditions were based on a 2 x 2 factorial design of Training by Elicitation. Levels of training were No Training and Probability Training. Levels of Elicitation were a Control group (independent forecasters), and Teams who worked in groups of 12 and were asked to justify their forecasts to each other. GJP had 3 additional conditions that were different forms of Prediction Markets. Finally, there was a condition of Superforecasters from Years 1 and 2.

Year 4 had 20 different conditions and approximately 13,000 forecasters. GJP devised three accountability conditions referred to as Outcome, Process, and Hybrid. Outcome accountability occurs when people were held accountable for making accurate predictions about the outcomes of events. Process accountability occurs when people are held accountable for the process they used to make a judgment or decision. In this condition, forecasters made comments that would help other forecasters make their predictions. (These forecasters were also asked to make numerical forecasts, just as those in the Outcome condition were asked to make comments that would help other forecasters improve their accuracy.) Comments were rated on the basis of “usefulness,” and forecasters in this condition were given feedback scores reflecting “usefulness.” Finally, hybrid accountability refers to situations in which people are held accountable for both accurate predictions and the process by which those predictions are made. Hybrid forecasters made both comments and predictions, and they received feedback based on both factors.

Year 4 included a variety of other studies. For example, GJP tested of the effect of team size in the survey, varying team size and had 1, 5, or 15 members. GJP also tested the effectiveness of batch auctions to set starting prices in the prediction markets. Although the introduction of the batch auction to determine prices on the first day did not increase the accuracy of the market’s predictions, it did reduce the skewness of total earnings. Finally, GJP examined 3 different conditions in which forecasters were asked to provide continuous probability distributions for the outcomes of each question.

II. Formal description of method for calculating rounding errors

As described in the paper’s “Data and Methods” section, our primary method for evaluating predictive accuracy is the commonly-used Brier Score. Brier Scores measure mean squared errors across assessments within a forecasting problem. Brier Scores are a function of predicted probabilities ( and observed outcomes (), where takes the value of 1 when outcome occurs and 0 when it does not. Brier Scores measure mean squared errors across assessments within a forecasting problem. The formula is , where is the number of predicted outcomes. As the Brier Score measures degrees of error, lower scores indicate more accurate assessments.

We adopted a deliberately conservative approach to our statistical analysis by treating the forecasting question as our unit of analysis. To do so, we first identify a subset of forecasters to evaluate (all forecasters, superforecasters, etc.). We then calculate an *aggregate Brier Score* for that group on each forecasting question. We calculate this score using the formula , where is a subset of GJP forecasters; is a forecasting question; is the mean of a vector; is a forecaster; is a day in the forecasting tournament; is the set of all forecasts made by forecaster on question while the question remained open;[[2]](#footnote-2) and is the Brier Score for an estimate made by a given forecaster on a given question on a given day. Thus, provides a question-level point-estimate of forecast accuracy among forecasters on question *j*.

We calculate *rounding errors* on forecasting questions by measuring proportional changes in Brier Scores when we round individual forecasts into bins of different widths. Thus, we define as a question-level point-estimate of forecasting accuracy among forecasters in group on question , having rounded individual respondents’ forecasts to the midpoints of bins. For example, we estimate the rounding error associated with transforming probabilities into three-step “confidence levels” as . Our findings also hold when we round predictions to the empirical expected value (that is, the frequency-weighted mean) of forecasts falling within each bin.

III. Full results for analysis of returns to precision across time horizons

As described in the main text, we divided forecasts into the following categories: (a) “Lay-Ups,” comprising forecasts made within two weeks of a question’s closing date and with no more than five percent probability or no less than ninety-five percent probability; (b) “Period I” forecasts made up to 36 days prior to a question’s closing date, but excluding Lay-Ups; (c) “Period II” forecasts made anywhere from 37 to 96 days prior to a question’s closing date; and (d) “Period III” forecasts made more than 96 days prior to a question’s closing date. There were 109,240 Lay-Ups in our data set, and 259,696 forecasts in the other three categories, respectively. Table A-1 analyzes returns to precision within each of these categories. Note that forecasters receive larger rounding penalties for Period II forecasts than for Period I forecasts: this is because we removed “Lay-Ups” from the Period I forecasts whereas Period II forecasts retain substantial numbers of small-probability estimates.

**Table A-1. Returns to precision across time horizons**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *Reference*  *class* |  | *Brier Score across all numerical forecasts* | *Rounding Errors* | | | | |
| 7 WEPs,  2015 version | 7 WEPs,  evenly spaced | Confidence levels  (3 bins) | Estimative verbs  (2 bins) | |
| *Lay-Ups* | | | | | | |
| All forecasters | *Mean:*  *Median:* | 0.065  0.009 | 42.4%\*\*\*  2.3%\*\*\* | 436.4%\*  40.4\*\*\* | 2578.1%\*\*  267.9%\*\*\* | 5865.1%\*\*\*  628.3%\*\*\* |
| Super-forecasters | *Mean:*  *Median:* | 0.046†  2.4e-4 | 869.4%  278.9\*\*\* | 5334.5%  2047.9%\*\*\* | 2.9e4%\*\*\*  1.2e4%\*\*\* | 6.6e4%\*\*\*  2.6e4%\*\*\* |
| *Period I forecasts: forecasts registered up to 36 days prior to question closing* | | | | | | |
| All forecasters | *Mean:*  *Median:* | 0.114  0.064 | 1.6%\*\*\*  1.2%\*\*\* | 8.1%\*\*\*  4.3%\* | 45.7%\*\*\*  15.8%\*\*\* | 118.0%\*\*\*  46.9%\*\*\* |
| Super-forecasters | *Mean:*  *Median:* | 0.073†  0.006 | 4.41%\*\*\*  6.9%\*\* | 29.1%\*\*\*  68.5%\*\*\* | 174.0%\*\*\*  432.2%\*\*\* | 422.7%\*\*\*  1032.5%\*\*\* |
| *Period II forecasts: forecasts registered between 37-96 days prior to question closing* | | | | | | |
| All forecasters | *Mean:*  *Median:* | 0.148  0.104 | 1.0%\*\*\*  1.0%\*\*\* | 3.8%  1.9%\* | 21.0%\*\*\*  8.5%\*\*\* | 55.6%\*\*\*  22.0%\*\*\* |
| Super-forecasters | *Mean:*  *Median:* | 0.119  0.047 | 32.0%\*\*\*  2.5%\*\* | 215.2%\*  14.5%\*\*\* | 1251.6%\*\*\*  97.3%\*\*\* | 2869.9%\*\*\*  255.1%\*\*\* |
| *Period III forecasts: forecasts registered at least 97 days prior to question closing* | | | | | | |
| All forecasters | *Mean:*  *Median:* | 0.187  0.155 | 0.7%\*\*\*  0.9%\*\*\* | 1.2%  0.8%\* | 6.6%\*\*\*  3.7%\*\*\* | 18.3%\*\*\*  11.8%\*\*\* |
| Super-forecasters | *Mean:*  *Median:* | 0.119  0.047 | 0.9%\*\*\*  0.7%\*\* | 13.5%\*\*  6.3%\*\*\* | 78.2%\*\*\*  22.6%\*\*\* | 204.4%\*\*\*  72.9%\*\*\* |

\*p<0.05, \*\* p<0.01, \*\*\* p<0.001. †On some questions, superforecasters’ average prediction was zero, creating a negatively-infinite rounding penalty. We dropped those observations in order to create the estimates shown here.

IV. Analysis of returns to precision across question types

To analyze whether returns to precision correlate with special clusters of questions, we used GJP’s data classifying forecasting questions with respect to 11 “region” tags and 15 “function” tags. The region tags corresponded to Sub-Saharan Africa, Central/South America, North America, South/Central Asia, East/Southeast Asia, Eastern Europe, Western Europe, Middle East/North Africa, Oceania, Global, and the Arctic. The function tags were Commodities, Currencies, Diplomatic Relations, Domestic Conflict, Economic Growth/Policy, Elections, International Organizations, International Security/Conflict, Leader Entry/Exit, Public Health, Resources/Environment, Technology, Trade, Treaties/Agreements, and Weapons. Tags were not mutually exclusive. Table A-2 describes the incidence of each regional and functional topic across GJP questions.

We examined these variables in ordinary least squares regressions predicting question-level thresholds based on dummy variables for question type. Table A-2 presents results. Model 1 combines all tags within a single regression. Models 2 and 3 examine region and function tags, respectively. Model 4 optimizes model fit, as measured by AIC score.

This purely inductive analysis is not intended to advance a theoretical framework for explaining question-level returns to precision. Rather, Table A-2 indicates the extent to which returns to precision belong to identifiable subsets of question types. And while all four models identify statistically significant patterns, none produces a particular high model fit. When we examine all 26 question tags simultaneously – in a regression that clearly entails overfitting – the model’s R2 is just 0.16. The constant term remains stable across models, indicating that baseline returns to precision are relatively unaffected when controlling for up to 26 question types.

Thus even in a purely inductive effort to identify how returns to precision vary across questions, we see little indication that forecasters’ ability to specify their estimates is confined to particular topics. These findings reinforce our broader argument that foreign policy analysts can consistently parse probabilities more finely than what common systems of qualitative expression allow.

**Table A-2. Predicting Question-Level Returns to Precision**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Model 1:  *All Tags* | Model 2:  *Region Tags* | Model 3:  *Function Tags* | Model 4:  *Optimal AIC* |
|  |  |  |  |  |
| Africa (*N*=34) | -0.88 (.99) | -0.91 (1.01) |  |  |
| Central/South America (18) | 1.02 (1.46) | 0.81 (1.54) |  | 1.31 (1.05) |
| North America (23) | 2.21 (1.20) | 1.64 (1.09) |  | 2.45 (1.00)\* |
| South/Central Asia (18) | 0.08 (1.30) | 0.18 (1.31) |  |  |
| East/Southeast Asia (80) | -0.38 (.85) | -0.30 (.86) |  |  |
| Eastern Europe (57) | -0.18 (.72) | -0.19 (.70) |  |  |
| Western Europe (56) | -1.36 (.75) | -1.86 (.76)\* |  | -1.02 (.63) |
| Global (6) | -2.02 (1.22)\*\* | -1.81 (1.40) |  |  |
| Mid. East/North Africa (125) | -0.59 (.81) | -0.17 (.77) |  |  |
| Oceania (5) | 4.89 (1.77) | 3.92 (2.77) |  | 4.71 (1.95)\* |
| Arctic (1) | -0.49 (1.63) | -2.07 (1.11) |  |  |
|  |  |  |  |  |
| Commodities (9) | -0.42 (1.24) |  | -0.94 (1.17) |  |
| Currencies (11) | -2.75 (1.21)\* |  | -1.84 (1.02) | -2.84 (1.37)\* |
| Diplomatic Relations (60) | 1.26 (.82) |  | 1.41 (.83) | 1.36 (.61)\* |
| Domestic Conflict (85) | -0.38 (.74) |  | -0.56 (.72) |  |
| Economic Growth (39) | -1.68 (.89) |  | -1.98 (.81)\* | -1.41 (.75) |
| Elections (49) | -2.21 (.73)\*\* |  | -2.44 (.71)\*\*\* | -2.27 (.67)\*\*\* |
| Int’l Organizations (33) | 0.81 (.97) |  | 0.50 (.90) |  |
| Int’l Security (75) | 0.41 (.70) |  | 0.31 (.70) |  |
| Leaders (47) | 1.39 (.96) |  | 1.44 (.95) | 1.45 (.69)\* |
| Public Health (3) | -2.57 (1.01)\* |  | -3.87 (1.02)\*\*\* | -3.34 (2.39) |
| Resources (17) | 0.39 (.98) |  | -0.06 (.89) |  |
| Technology (4) | 2.80 (1.11)\* |  | 2.16 (.61)\*\*\* |  |
| Trade (22) | -3.41 (.83)\*\*\* |  | -2.01 (.77)\*\* | -3.20 (.98)\*\*\* |
| Treaties (41) | -0.05 (.73) |  | -0.07 (.69) |  |
| Weapons (28) | -0.76 (.84) |  | -0.71 (.81) |  |
|  |  |  |  |  |
| Constant | 6.84 (1.15)\*\*\* | 6.48 (.74)\*\*\* | 6.53 (.70)\*\*\* | 6.35 (.31)\*\*\* |
|  |  |  |  |  |
| Observations | 375 | 375 | 375 | 375 |
| R2 | 0.16 | 0.05 | 0.11 | 0.15 |
| AIC | 2,153 | 2,169 | 2,156 | 2,132 |
|  |  |  |  |  |

Ordinary least squares regression predicting variation in question-level thresholds, with robust standard errors.

V. Additional description of the Good Judgment Project’s training module

As described above, a random subset of GJP forecasters received online training. Beginning on the following page, we have enclosed a copy of the GJP’s training manual, from which we have removed copyrighted images and embedded interactive elements.

Training Manual:

CHAMPS & KNOW for Surveys

(Copyrighted images and embedded interactive elements were removed from this version.)

Basic CHAMPS & KNOW for Surveys

**Lessons from past seasons**

* One key lesson from the past two seasons of the IARPA tournament is that good geopolitical forecasting is a matter of skill, not just luck.
* Good forecasting is a skill that you can cultivate through training and deliberate practice.
* Good Judgment Project forecasters who completed a probabilistic-reasoning training module in past seasons did significantly better in the tournament than those who did not.

**Skill and Principle**

In this season, we are building on the lessons from the past years. We will be providing you more training covering both probabilistic-reasoning **skills** and substantive **principles** of political analysis that you can apply to forecasting problems.

Lessons:

1) Training is so important that we require you to:

* Complete this overview training module before you access your forecasting website for the first time; and
* Complete two additional training modules during the “preseason” before scored forecasting begins in August 2013.

2) Learning does not stop here:

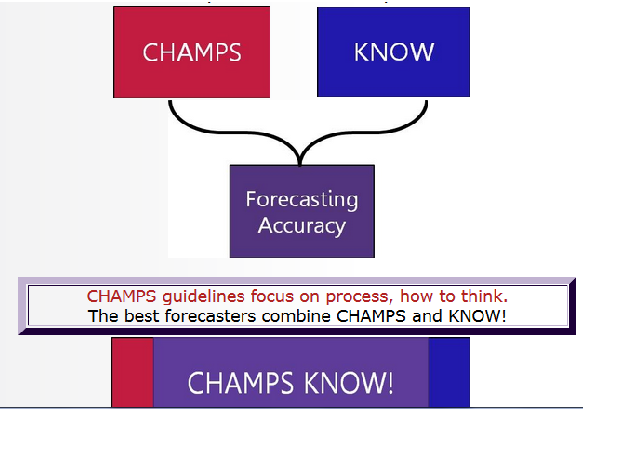
* Learn by applying principles to actual forecasting problems.
* Learn by taking optional advanced training later in the season.

3) Our goal is:

* To give you guidelines for improving your accuracy in Season 3 of the tournament; and
* To test your understanding of key concepts.

**What makes a good forecaster?**

Our training tips can be condensed into these acronyms:



**Introducing our guidelines**

* This overview module will give you a first look at the “CHAMPS KNOW” guidelines.
* Right now, we are introducing the concepts to help you recognize their possible uses when you first log in to your forecasting website and see your first Season 3 questions.
* We will point out features of our forecasting website that will help you apply these concepts.
* Later on you’ll review two in-depth training modules on the “CHAMPS” process guidelines and the “KNOW” content guidelines.

**Introducing CHAMPS**

**CHAMPS Guidelines**

What is CHAMPS... six key steps:

* **C**omparison classes inform probability estimates.
* **H**unt for the right information.
* **A**djust and update forecasts when appropriate.
* **M**athematical and statistical models can help.
* **P**ost-mortem analysis after outcomes are known.
* **S**elect the right questions to answer

**C**HAMPS

Comparison classes should inform your probability estimates

**CHAMPS: Comparison Classes**

**What are the chances that a married couple will get divorced?** The **outside view:** connect the case at hand to a comparison or reference class and rely on base-rate information.

* EXAMPLE: I know 40% of all marriages end in divorce. So, the wedding I just attended has a 40% chance of ending in divorce.

The **inside view:** focus on unique qualities of case at hand.

* EXAMPLE: The couple seems well matched. So, I assign a low probability that their marriage will end in divorce.

**Research suggests that people over-RELY on the inside view!**

“The inside view is overwhelmingly preferred in intuitive forecasting. The natural way to think about a problem is to bring to bear all one knows about it, with special attention to its unique features. The intellectual detour into the statistics of related cases is seldom chosen spontaneously”

~Daniel Kahneman,

An advisor to the Good Judgment Project

**Exercise 1:** The Linda problem  
  
Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.  
  
Which of the following two alternatives is more probable?

* 1. Linda is a bank teller
  2. Linda is a bank teller and active in the feminist movement

**ANSWER:** There must be at least as many bank tellers as there are feminist bank tellers, and as such, it is more probable that Linda is a bank teller than a feminist bank teller.

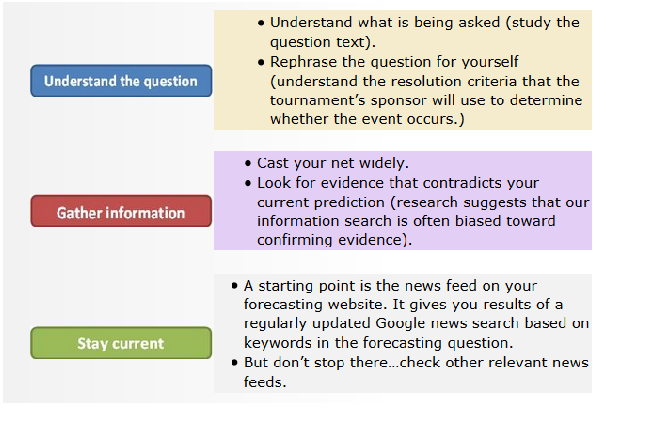
**CHAMPS: Comparison Classes**

* To apply the outside view, you need to identify an appropriate **base rate**.
* Base rates should be drawn from **comparison classes** most relevant to the forecast at hand.
  + A comparison class is a set whose members (people, nations, etc.) share key attributes.
  + The base rate is the proportion of target events in a class of events under consideration.

EXAMPLE: One base rate for the risk of divorce over a one-year period could be calculated by looking up the number of divorces that occurred in a recent year, and dividing that by the total number of married couples at the beginning of that year.

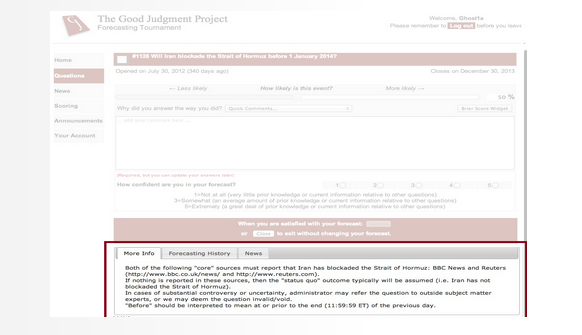
**CHAMPS – Hunt for the right information**

**To make accurate predictions, you need to:**

****

**CHAMPS: More Info**

"More Info" feature on the forecasting website (platform) gives you the resolution criteria.



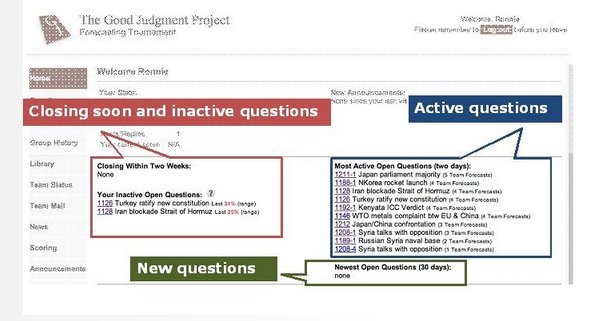
**CHAMPS – Adjust and update forecasts when appropriate**

**CHAMPS: Adjust often**

* Adjusting your forecasts as events dictate is another key to forecasting success.
* Forecasting is not just about guessing once and walking away. It is about updating and converging on the truth as quickly as possible.
* Your predictions should reflect your best estimate at that moment. Stale forecasts will hurt your Brier score.
* Our best forecasters regularly update their forecasts to take account of:
  + New information
  + Their changing assessments of prior information
  + The passage of time

**CHAMPS: Platform**

The forecasting platform can help you to identify questions that you should revisit for possible updating.

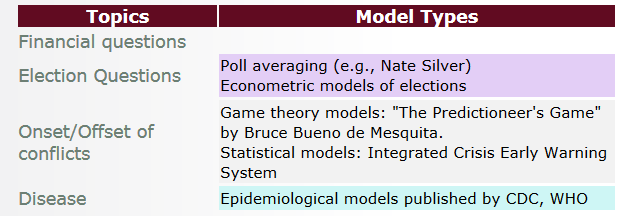


**CHAMPS – Mathematical and statistical models can sometimes be helpful**

**CHAMPS: Mathematical and statistical models**

* When models with good track records are available, exploit them.
* Many different models can help you make a better forecast:
  + Simple averaging (e.g., mean, median, mode)
  + Weighted averaging (e.g., higher weights to recent forecasts)
  + Domain-specific models (e.g., use only forecasters who have answered similar events correctly)

**Be relentlessly opportunistic in seeking out useful models**



**CHAMPS – Post-mortem analysis after outcomes are known**

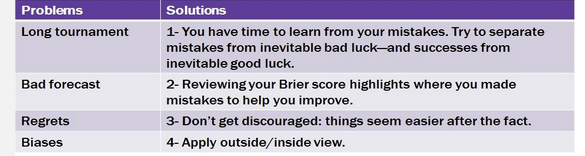
**CHAMPS – Post-mortems**

One of the greatest tension is to draw a narrow or broad conclusion from post-mortem analysis.

Ask yourself if your mistakes are from uncertainty (just bad luck) or a product of your forecasting strategy (a failure to apply the principles of training, human error). If it is uncertainty then don't over think it.

**CHAMPS – Post-mortem analysis**

The following describes problems from post-mortem analysis. Each one is matched to its proper solution:



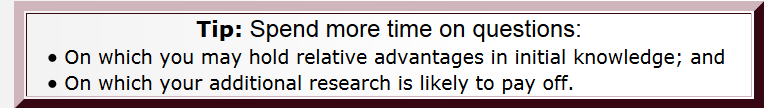
**Be careful about how much you adjust your strategy**

Over-reaction can hurt you, too!!

**The mirror-image mistake:** When we catch ourselves making an error in one direction, we often over-compensate and try the exact opposite strategy.

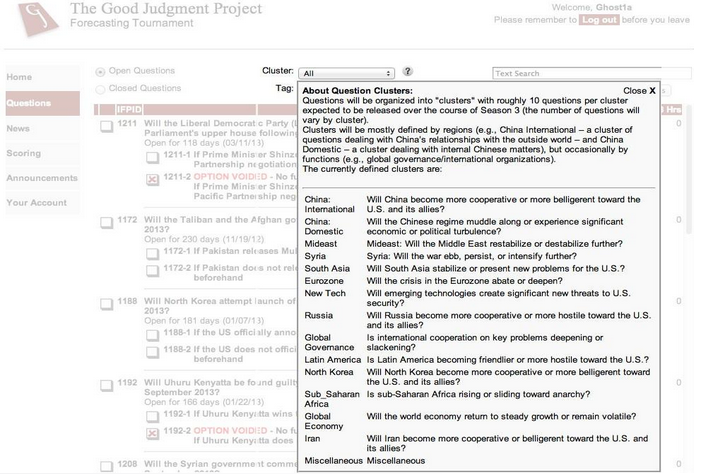
**CHAMPS: Select the right questions**

* Selecting the right questions may be as important as doing your best to answer them.
* Just as physicians have to choose how to allocate attention across patients, you will have to choose how to allocate attention across questions.



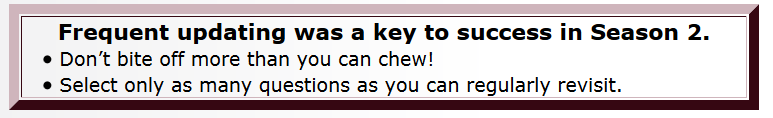
**CHAMPS: Question clusters**

Individual Forecasting Problems (“IFPs” or questions) will be organized into "clusters" with roughly 10 questions per cluster. These clusters will help you select the right IFP by region and sometimes by function.



**Be careful about which questions you take on**

* We expect 150 forecasting problems in Season 3, and few forecasters will be able to answer all well.
* Scoring encourages you to select challenging questions on which you have a comparative advantage.
* This means selecting some hard questions, not just easy ones (hard to beat the crowd on just easy questions).



**INTRODUCING KNOW**

**KNOW Guidelines**

What is KNOW... four key insights:

* **K**now the power players and their preferences.
* **N**orms and protocols of domestic and international institutions matter.
* **O**ther perspectives (bottom-up eruptions of mass protest, globalization/interconnected-world arguments, efficient-markets thesis and its critics…) should also inform your forecasts.
* **W**ildcards, accidents and black swans can catch you off-guard. Beware of irreducible uncertainty.

**KNOW: Know the Power Players**

Here is one way to structure your inside-the-case view analysis: **POM**

* **Players:** Identify political players and their relative power.
* **Objectives:** Estimate what each wants on this issue.
* **Motivation:** Estimate how motivated each is to prevail (spend political capital to achieve preferred outcomes).

Link this inside-view to an outside-view base rate. All else being equal, many experts believe that power players motivated to prevail on an issue will win (outside-view, base-rate prediction)—unless their influence is offset by equally powerful players motivated to oppose them.

**KNOW: Norms and protocols of domestic and international institutions**

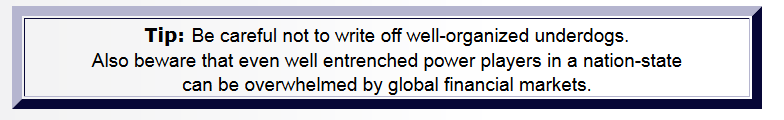
One set of qualifiers on the power-politics perspective. For example:

* Understanding institutional ground rules–of courts, parliaments and international organizations–can be crucial to forecasting success.
* Such rules can amplify the influence of some power players—and attenuate the influence of others:
* **Power-enhancing rules:** e.g., the U.N. Security Council veto enhances the power of the five permanent members relative to other nations.
* **Power-diminishing rules:** e.g., the U.S. Senate filibuster rules diminish the power of the majority power.

**KNOW: Other perspectives, in addition to power politics and institutions**

Other perspectives can take diverse forms. For example:

* Revolutionary leaders—Gandhi, Khomeini and Mandela—have shown that power does NOT always flow top-down.
* Idealists/fanatics sometimes defy power-political logic (although betting on underdogs is risky—especially in short time frames).
* Some experts argue that mass mobilization against oppressive regimes has grown easier as the world has grown interconnected.
* And other experts argue that the global markets for capital and labor are increasingly efficient, making it harder for isolated regimes to survive.



**KNOW: Wildcards, accidents and black swans**

***Black Swans?***

“A black swan is an extreme-impact and extreme-outlier event that virtually no one expected but that, after the fact, many see as inevitable (hindsight bias)”

~Nassim Taleb,

An advisor to the Good Judgment Project

**KNOW: Wildcards, accidents and black swans**

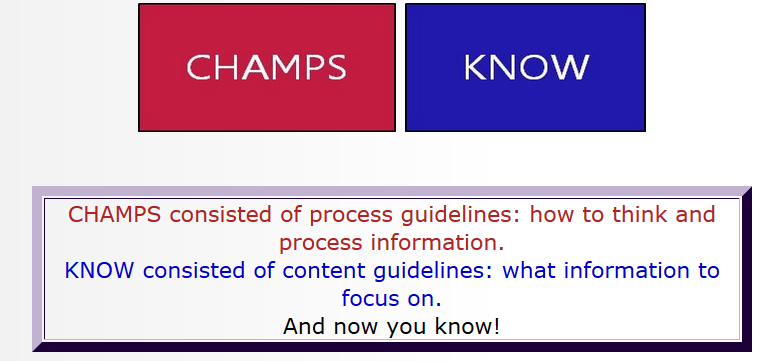
Even when you are right about the power-politics and the institutional ground rules, you can get it wrong because…

* Low-level actors can move events against the wishes of those formally in charge (e.g., fishing boat captains in East China Sea).
* Accidents and disasters happen: bombs hit unintended targets, tsunamis stun economies and suspend wars.
* Unknown unknowns can cause rare but important events.
* All questions can be affected by fluky forces, but some are more vulnerable to others.
* Questions that may pivot on small-scale interactions outside the direct control of power players tend to be higher in irreducible uncertainty. So too are questions that pivot on potentially volatile financial markets.

**CHAMS KNOW Summary**

We have given you a quick overview of the principles embodied in

CHAMPS and KNOW



**Quick Reference**

**C**omparison classes should inform your probability estimates.

**H**unt for the right information.

**A**djust and update your forecasts when appropriate.

**M**athematical and statistical models can help.

**P**ost-mortem analyses help you improve.

**S**elect the right questions to answer.

**K**now the power players.

**N**orms and protocols of domestic and international institutions matter.

**O**ther perspectives aside from power politics can also inform your forecasts.

**W**ildcards, accidents and black swans can catch you off-guard if you don’t consider the risk of irreducible uncertainty.

VI. Descriptive statistics for individual attributes

Table A-3 summarizes all variables used to examine variations in individual-level returns to precision in the penultimate section of the paper.

**Table A-3. Summary Statistics for Individual-Level Attributes**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *N* | *Mean* | *Std. dev.* | *Min* | *Max* |
| *Returns to Precision* |  |  |  |  |  |
| Threshold of Estimative Precision () | 1,832 | 4.36 | 8.03 | 1.00 | 101.00 |
|  |  |  |  |  |  |
| *Targets for Cultivation* |  |  |  |  |  |
| Median Brier Score | 1,832 | 0.06 | 0.06 | 0.00 | 1.00 |
| Number of Questions | 1,832 | 85.13 | 62.92 | 25.00 | 375.00 |
| Average Revisions per Question | 1,832 | 2.49 | 5.03 | 1.00 | 101.40 |
| Granularity | 1,832 | 0.50 | 0.19 | 0.00 | 1.00 |
| Probabilistic Training | 1,832 | 0.65 | 0.48 | 0.00 | 1.00 |
| Group Collaboration | 1,832 | 0.51 | 0.50 | 0.00 | 1.00 |
|  | | | | |  |
| *Targets for Selection* | | | | |  |
| Education Level | 1,821 | 1.78 | 0.79 | 0.00 | 3.00 |
| Numeracy | 1,817 | 0.05 | 0.91 | -4.80 | 1.11 |
| Raven’s Progressive Matrices | 1,822 | 7.67 | 2.72 | 0.00 | 12.00 |
| Cognitive Reflection Test | 1,620 | -0.04 | 0.78 | -3.41 | 1.11 |
| Fox-Hedgehog | 1,822 | 1.86 | 1.15 | 0.00 | 5.00 |
| Need for Cognition | 1,822 | 5.05 | 1.86 | 0.00 | 7.00 |
|  |  |  |  |  |  |
| *Additional Controls* |  |  |  |  |  |
| Age | 1,818 | 39.58 | 13.21 | 20.00 | 91.00 |
| Female | 1,826 | 0.17 | 0.38 | 0.00 | 1.00 |
| Superforecaster | 1,832 | 0.04 | 0.20 | 0.00 | 1.00 |
| Twenty respondents’ thresholds were Winsorized to 21 bins (i.e., the level of precision corresponding to increments of five percentage points) when estimating bivariate correlations, so as to reduce the impact of outliers. | | | | | |

1. Here we provide information on the Good Judgment Project as a whole. Note that since we only used a subset of these data in our analysis (see paper for details) some of these descriptive statistics differ from what is presented in the manuscript. [↑](#footnote-ref-1)
2. With a maximum of one forecast per day, recorded as a forecaster’s most recent estimate prior to midnight, U.S. Eastern Time. [↑](#footnote-ref-2)