Challenges and developments in Structured Expert Judgement

Prof Tim Bedford
University of Strathclyde
COST IS1304 Action on Structured Expert Judgement

Introduction

- Major objective of this Action is to be able to encourage senior policy/DMs to use SEJ
- Discussions indicate
  - awareness of EJ, low understanding of SEJ
  - Some awareness of different approaches
- Academic literature
  - Much work on EJ/SEJ from different disciplines
  - Entrenched positions create confusion in users
  - Until recently, limited empirical research
  - Limited attempts to incorporate contextual issues into selection of appropriate methods
- Diversity of methods available, some attracting $$$

Expert Judgement approaches

- Delphi – developed after WW2 by RAND, disavowed, and rehabilitated
- Nominal Group Technique
- Stanford Research Institute Process
- NUREG
- Psychological Scaling Techniques
- Classical Model
- SHELF
- Prediction markets
- Superforecasters – IARPA ACE competition

Questions being asked…(broadly)..

- Scoping
- Simplifying
- Predicting
- Deciding

The Decision Makers job, not the experts job, or the analysts

Context

- Considering *predictions* area, can we usefully define different contextual factors that would allow us to differentiate between “good practice” SEJ approaches?

Some important contextual issues

- Extent to which (standard) modelling approach(es) and/or data exists and is relevant
- Speed of application
- Many experts available or highly specialised
- Societal accountability (eg private company/public authority)
- Game-playing, adversarial and other behavioural responses
- Consensus- validation,onside, speed

Some important c s

Understanding

Speed of application

Legitimation burden

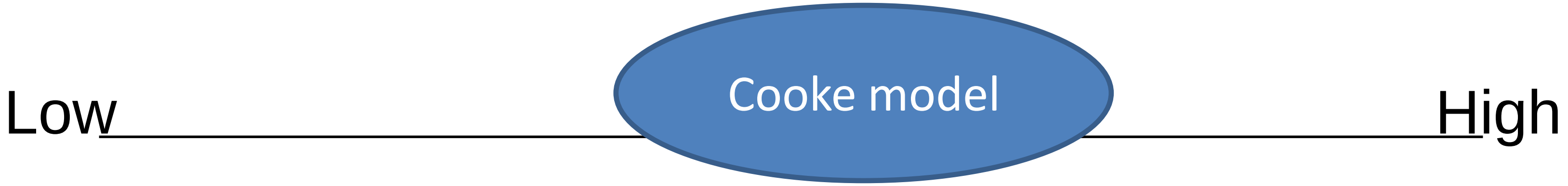
- Extent to which (standard) meth pproach(es) and/or data exists and is re
- Speed of application
- Many experts available or highly specialised
- Societal accountability (eg private company/public authority)
- Game-playing, adversarial and other behavioural responses
- Consensus- valid peed

Understanding

Legitimation burden

Speed of application

Degree of understanding



Low

Cooke model

High

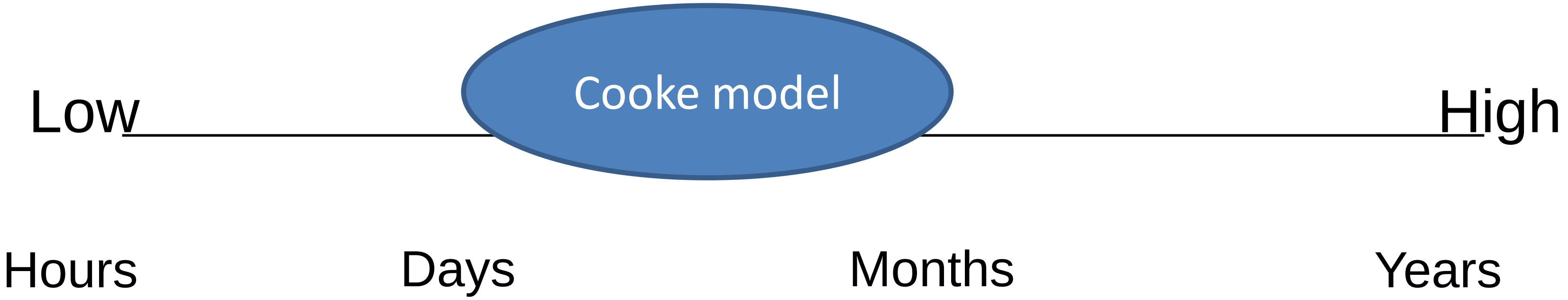
Lack of relevant data or models with explanatory value

Competing models with explanatory value

Models with explanatory value and some relevant empirical data

Excellent explanatory models and relevant empirical data, giving good predictive power in relevant contexts

Time available for application



Cooke model

Low

High

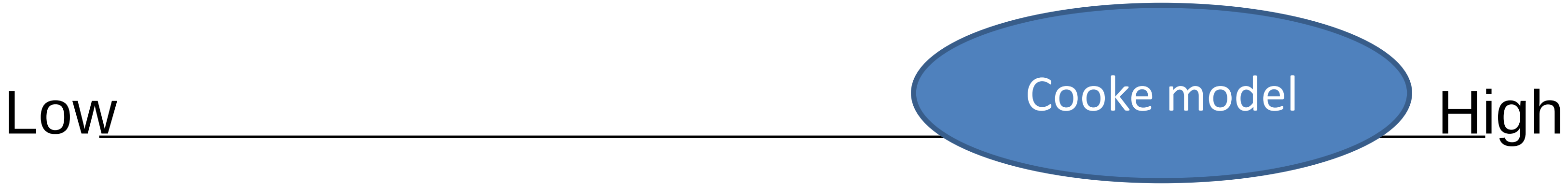
Hours

Days

Months

Years

Legitimation burden



Low

Cooke model

High

Internal expertise, small numbers of experts with an interest in outcome and no external validation

Consensus driven, but with experts who have no interest in outcome

External validation and quality process but small number of experts

External validation and evidence of quality of the process and validators

Sheffield Elicitation Framework (SHELF)

- Uses behavioural aggregation: expert group is asked to collectively agree to a distribution that a Rational Impartial Observer (RIO) would agree to
- O’Hagan strongly believes that this makes more sense than weighting experts and taking mixture distribution (cf Classical Model)
- 2-5 distributions assessed in 1-2 day workshop

SHELF – 2

- Discussion
- Training
- Individual assessment
- Group discussion about individual assessments
- Agree group consensus
- Fit a distribution using software

Experts may not actually agree, so may have to agree to differ. Unclear what implications are in practice.

Note distinction between what the expert thinks and what they agree a RIO might agree

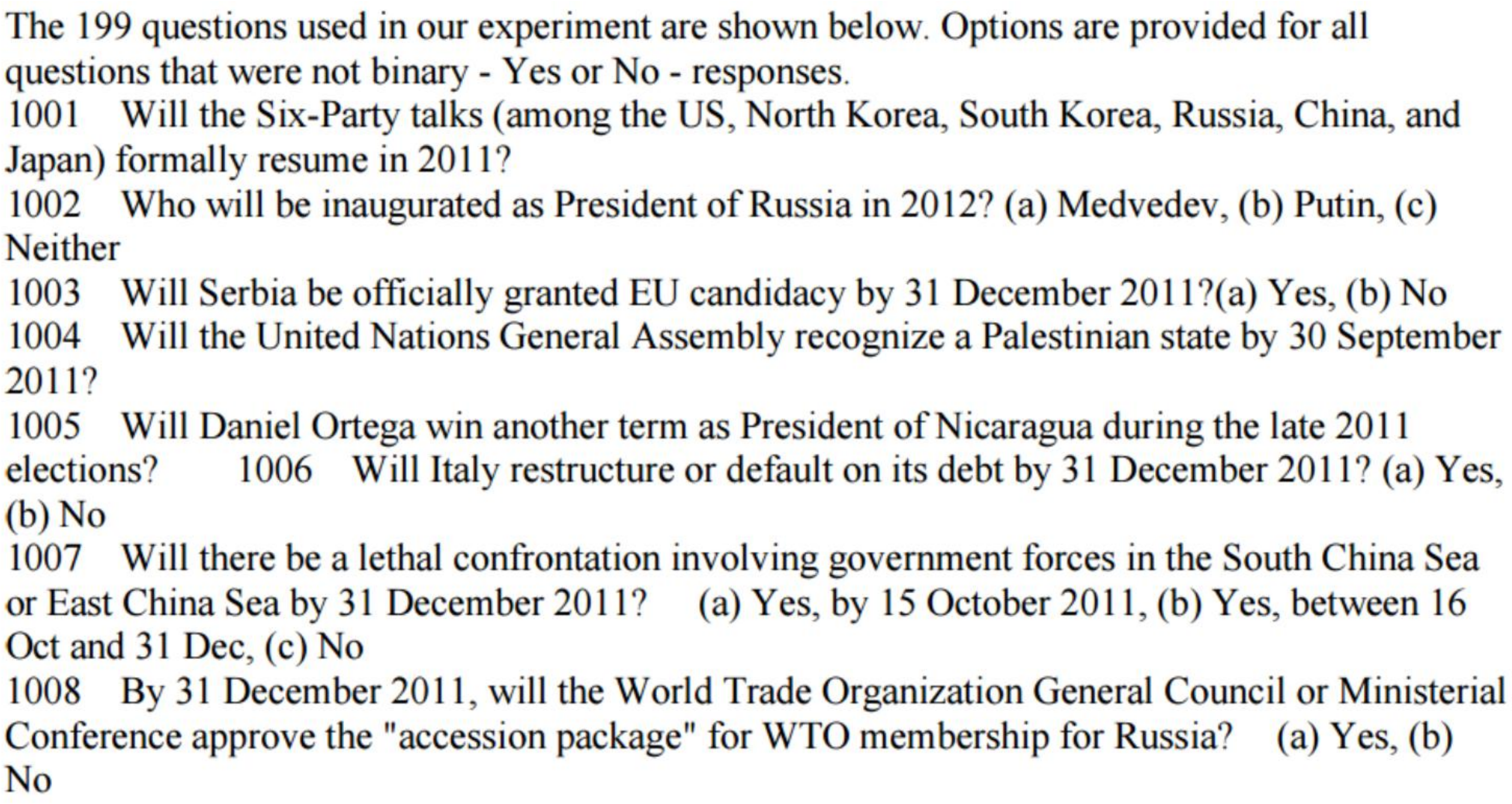
Comparison with Classical Model

- No calibration of experts… if there is any data then this is fed to the experts so that they can learn or take account of this
- “Ideological” difference about the meaning of a weighted mixture distribution
- Process of elicitation has to deal with all biases etc
- Process of discussion between experts is similar to other methods.
- *Question: is overall result better when you let experts learn from seed questions, or when you use seed questions to down-weight poorly performing experts?*

IARPA

- IARPA - Intelligence Advanced Research Projects Activity
- *Aggregative Contingent Estimation Program* run by *Office for Anticipating Surprise*
- Prediction tournament included 5 academic teams able to test different methods over 2011-14
- Focus on geopolitical uncertainties
- Tournament won by *Good Judgement Project*, now operating commercially and “open”.

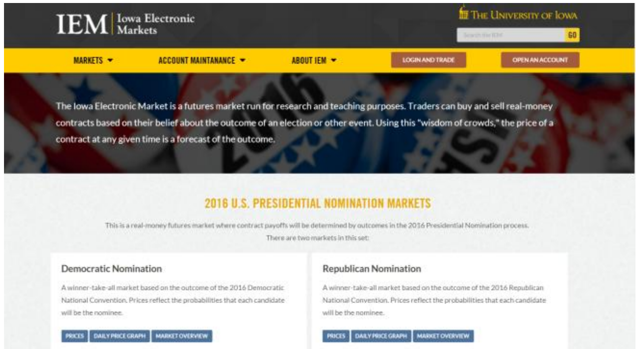
Example IARPA/GJP questions



The 199 questions used in our experiment are shown below. Options are provided for all
questions that were not binary - Yes or No - responses.
1001 Will the Six-Party talks (among the US, North Korea, South Korea, Russia, China, and
Japan) formally resume in 2011?
1002 Who will be inaugurated as President of Russia in 2012? (a) Medvedev, (b) Putin, (c)
Neither
1003 Will Serbia be officially granted EU candidacy by 31 December 2011?(a) Yes, (b) No
1004 Will the United Nations General Assembly recognize a Palestinian state by 30 September
2011?
1005 Will Daniel Ortega win another term as President of Nicaragua during the late 2011
elections? 1006 Will Italy restructure or default on its debt by 31 December 2011? (a) Yes,
(b) No
1007 Will there be a lethal confrontation involving government forces in the South China Sea
or East China Sea by 31 December 2011? (a) Yes, by 15 October 2011, (b) Yes, between 16
Oct and 31 Dec, (c) No
1008 By 31 December 2011, will the World Trade Organization General Council or Ministerial
Conference approve the "accession package" for WTO membership for Russia? (a) Yes, (b)
No

Prediction Markets

- Uses trades in an electronic money market to provide an indication of probabilities
- Both commercial and academic/research sites exist
- Participants buy futures in outcomes, eg Trump wins the presidency
- Eg Future pays $1 if Trump wins, and nothing otherwise. These futures are traded, and you can buy, sell etc
- Iowa Electronic Market has permission to trade from the US authorities – limited stakes/winnings – as online gambling is illegal in the US
- Consensus Point provides commercial “crowdsourced” advice



The Iowa Electronic Market is a futures market run for research and teaching purposes. Traders can buy and sell real-money contracts based on their belief about the outcome of an election or other event. Using this "wisdom of crowds," the price of a contract at any given time is a forecast of the outcome.

2016 U.S. PRESIDENTIAL NOMINATION MARKETS

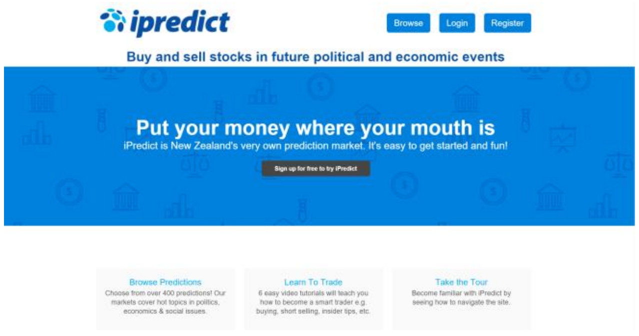
This is a real-money futures market where contract payoffs will be determined by outcomes in the 2016 Presidential Nomination process. There are two markets in this set:

Democratic Nomination

A winner-take-all market based on the outcome of the 2016 Democratic National Convention. Prices reflect the probabilities that each candidate will be the nominee.

Republican Nomination

A winner-take-all market based on the outcome of the 2016 Republican National Convention. Prices reflect the probabilities that each candidate will be the nominee.



Buy and sell stocks in future political and economic events

Put your money where your mouth is

iPredict is New Zealand's very own prediction market. It's easy to get started and fun!

Browse Predictions

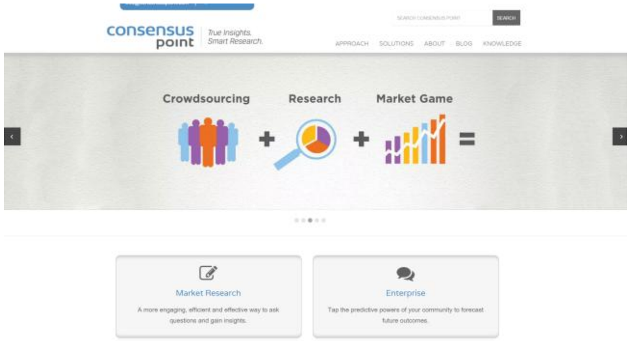
Choose from over 400 predictions! Our markets cover hot topics in politics, economics & social issues.

Learn To Trade

6 easy video tutorials will teach you how to become a smart trader e.g. buying, short selling, insider tips, etc.

Take the Tour

Become familiar with iPredict by seeing how to navigate the site.



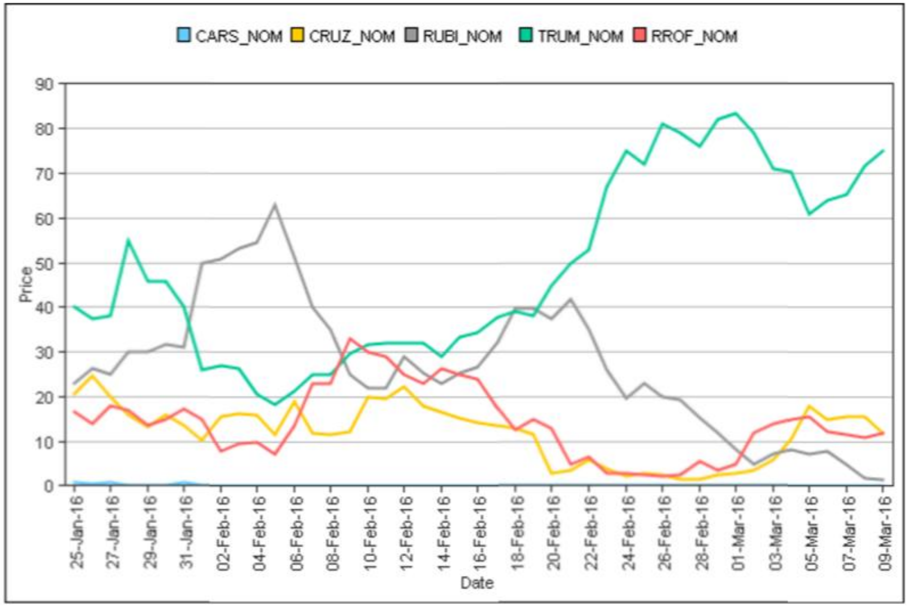
Market Research

A more engaging, efficient and effective way to ask questions and gain insights.

Enterprise

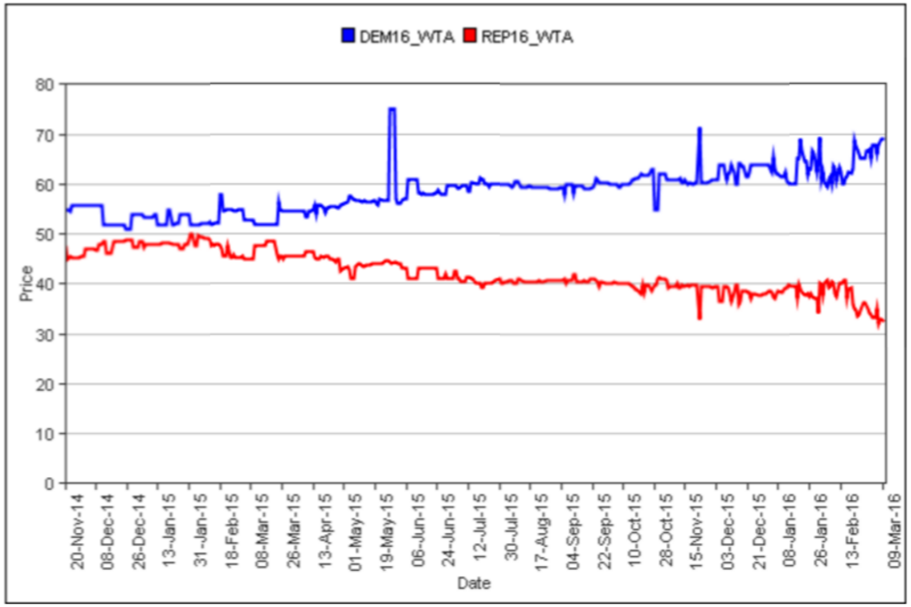
Tap the predictive powers of your community to forecast future outcomes.

Iowa current prices for the US Presidential election

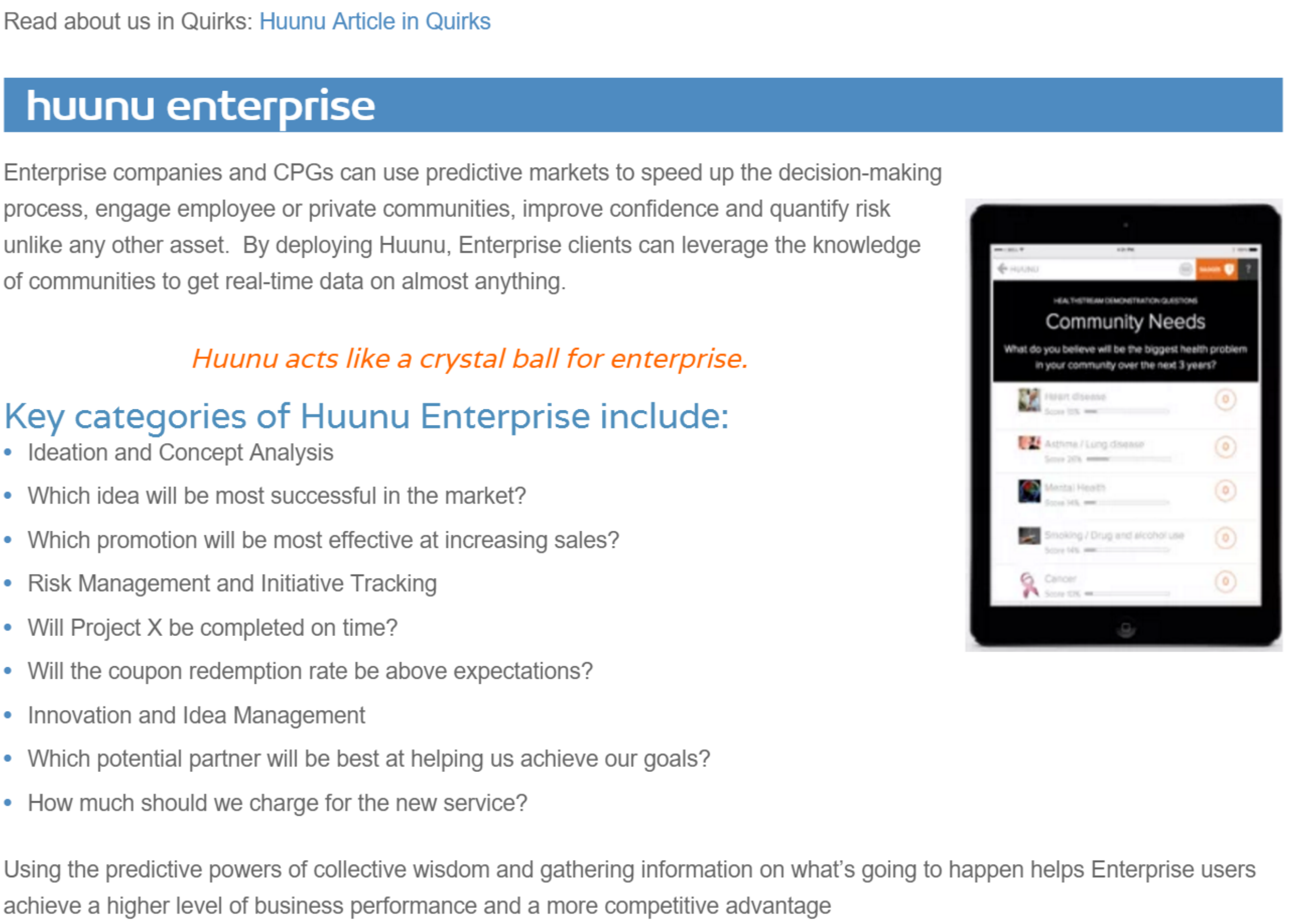


Who will be the Republican nominee?

Will the Democrats or the Republicans win?



Commercial applications…



Read about us in Quirks: Huunu Article in Quirks

huunu enterprise

Enterprise companies and CPGs can use predictive markets to speed up the decision-making process, engage employee or private communities, improve confidence and quantify risk unlike any other asset. By deploying Huunu, Enterprise clients can leverage the knowledge of communities to get real-time data on almost anything.

Huunu acts like a crystal ball for enterprise.

Key categories of Huunu Enterprise include:

- Ideation and Concept Analysis
- Which idea will be most successful in the market?
- Which promotion will be most effective at increasing sales?
- Risk Management and Initiative Tracking
- Will Project X be completed on time?
- Will the coupon redemption rate be above expectations?
- Innovation and Idea Management
- Which potential partner will be best at helping us achieve our goals?
- How much should we charge for the new service?

Using the predictive powers of collective wisdom and gathering information on what's going to happen helps Enterprise users achieve a higher level of business performance and a more competitive advantage

HEALTHSTREAM DEMONSTRATION QUESTIONS

Community Needs

What do you believe will be the biggest health problem in your community over the next 3 years?

Heart disease
Score 15%

Asthma / Lung disease
Score 26%

Mental Health
Score 14%

Smoking / Drug and alcohol use
Score 14%

Cancer
Score 10%

Good Judgement Project

- Project led by Tetlock from U of Pennsylvania
- Recruited large numbers of potential experts to answer questions
- Different groups 3x4
  - Not trained, probability training, scenario training
  - Individual, Crowd-informed individuals, Interactive Group, Prediction Market
- After 1 year, created a Superforecasters group

GJP results

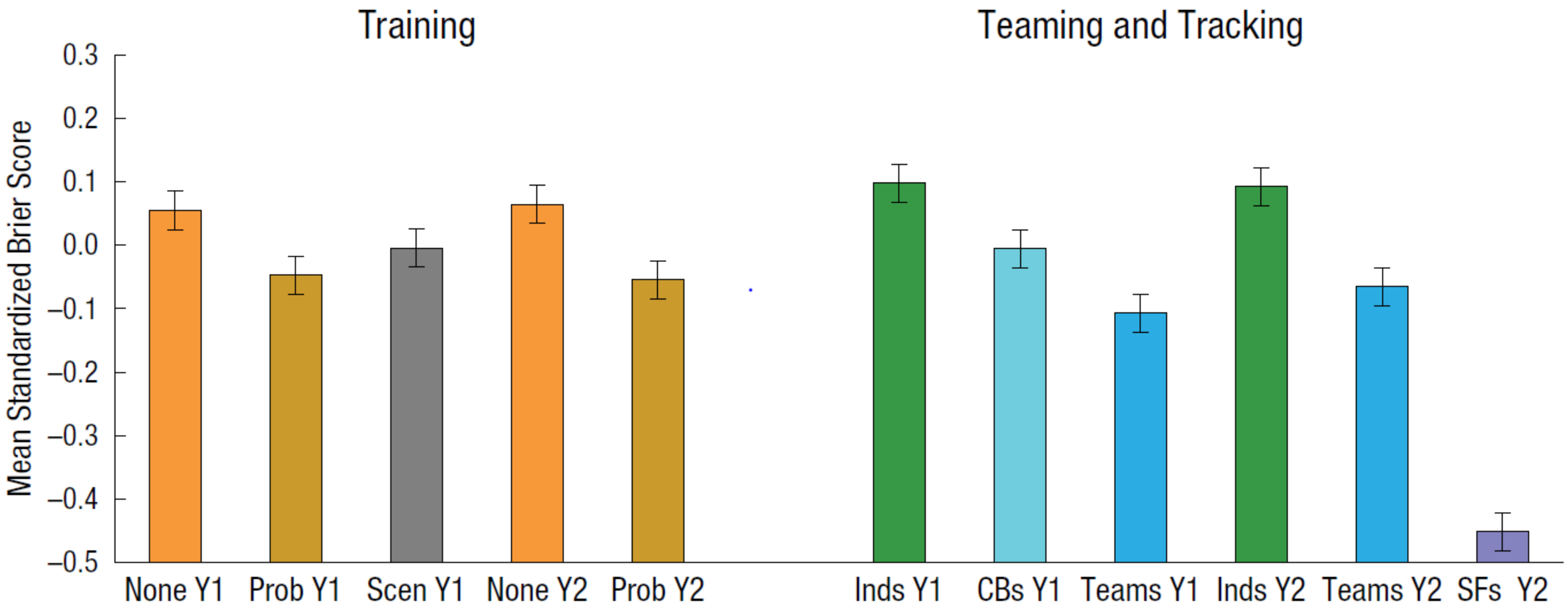
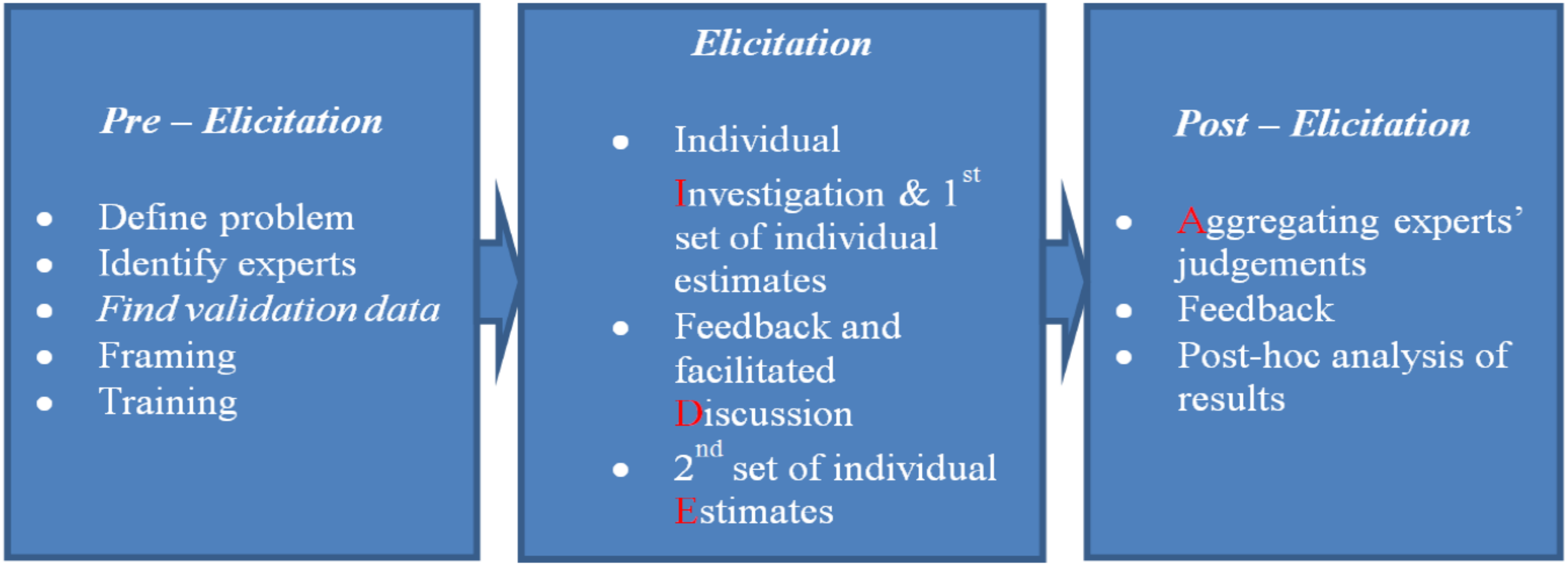


Fig. 1. Effects of training, teaming, and tracking on average Brier scores in Year 1 (Y1) and Year (Y2). The bars at the left show results for the no-training ("None"), probability-training ("Prob"), and scenario-training ("Scen") conditions; the bars at the right show results for independent forecasters ("Inds"), crowd-belief forecasters ("CBs"), team forecasters ("Teams"), and superforecasters ("SFs"). Error bars represent ±2 *SE*s.

B. Mellers, L. Ungar et al, Psychological Strategies for Winning a Geopolitical Forecasting Tournament, Psychological Science 2014, Vol. 25(5) 1106–1115

IDEA

- Due to Burgman et al – a mixed method



Pre – Elicitation

- Define problem
- Identify experts
- *Find validation data*
- Framing
- Training

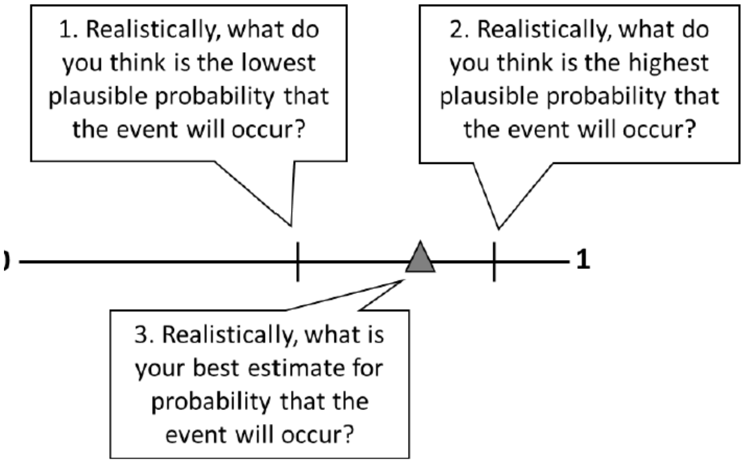
Elicitation

- Individual Investigation & 1st set of individual estimates
- Feedback and facilitated Discussion
- 2nd set of individual Estimates

Post – Elicitation

- Aggregating experts' judgements
- Feedback
- Post-hoc analysis of results

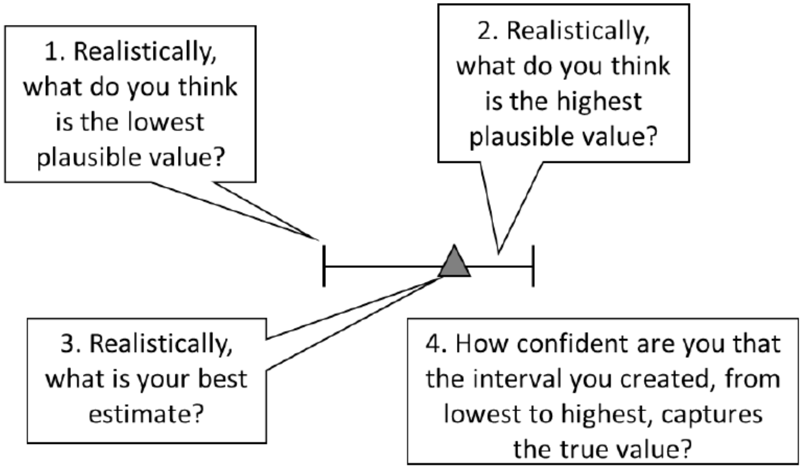
IDEA elicitation protocol



1. Realistically, what do you think is the lowest plausible probability that the event will occur?

2. Realistically, what do you think is the highest plausible probability that the event will occur?

3. Realistically, what is your best estimate for probability that the event will occur?



1. Realistically, what do you think is the lowest plausible value?

2. Realistically, what do you think is the highest plausible value?

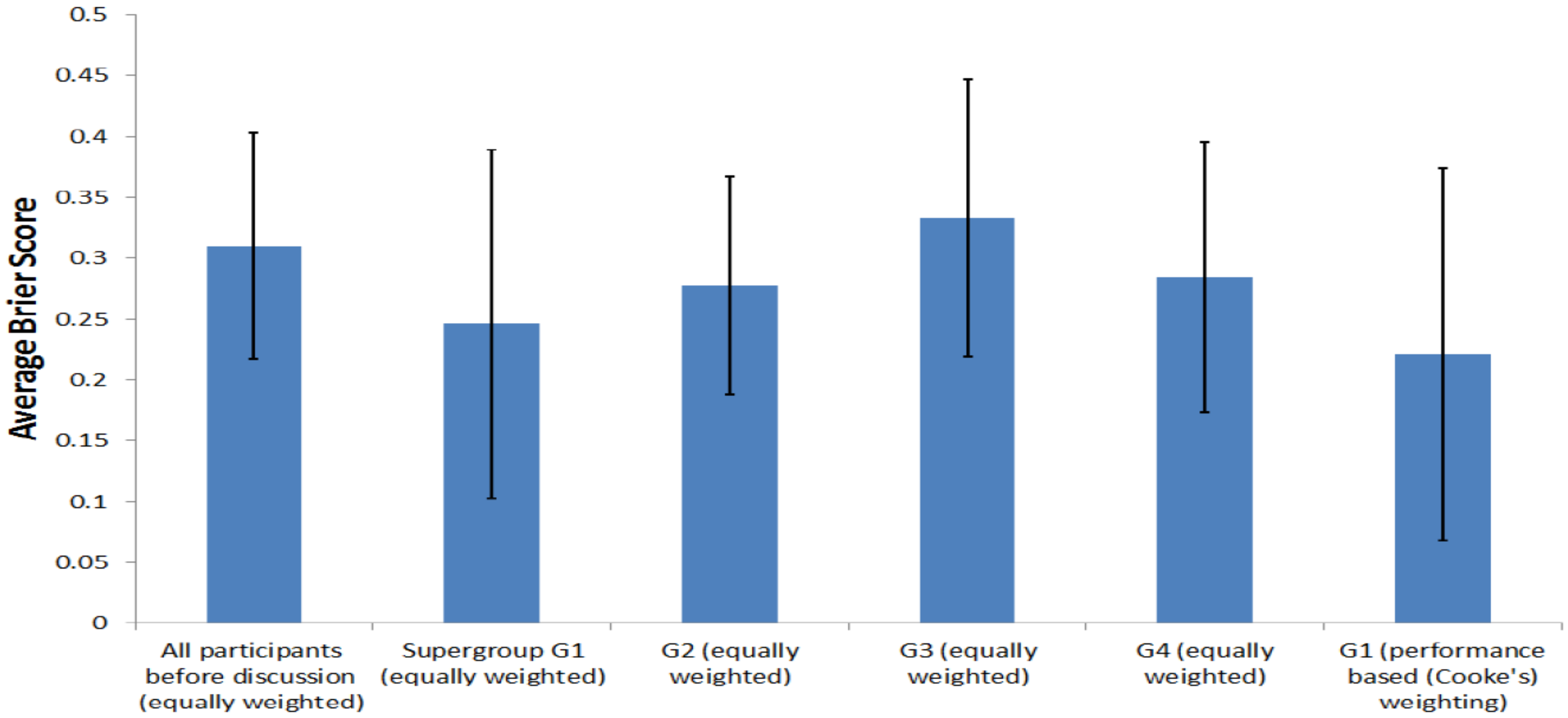
3. Realistically, what is your best estimate?

4. How confident are you that the interval you created, from lowest to highest, captures the true value?

For events

For variables

IDEA Year 2 performance in IARPA



Comparison of GJP and IDEA

- Size of expert groups (GJP>>IDEA)
- IDEA asks for ranges, GJP for point values
- Performance based weighting in IDEA
- More structured approach to facilitation in IDEA (possible with smaller group).
- GJP did better overall but IDEA was not far behind
- Cost of carrying out the protocols

Key considerations in designing a group elicitation process

- Expert interaction positives
  - Ensure all understand the questions and eliminate incorrect (narrowing) assumptions
  - Agree qualitative structure of the problem, hence simplifying the set of questions that need elicitation
  - Discussion about potential mechanisms, base rates, comparative classes etc, highlights aspects that should be considered
- Expert interaction negatives
  - Development of “groupthink” - Focus on one or two mechanisms, or comparative classes
  - Non-expertise based influences (eg ability to articulate, dominant personality, peer esteem, job level)

Conclusions

- Acceptance of SEJ increasing ($$$)
- Training helps a bit
- Weighting of experts helps a bit more
- Getting good experts together to discuss rationales helps a lot (identified by performance)
- Performance weighting still helps!

Brier score – a proper scoring rule

- Suppose expert provides probabilities $p_1, \dots, p_n$ for exclusive outcomes $1, \dots, n$.
- Define $x_i = \begin{cases} 1, & \text{if } i \text{ occurs} \\ 0, & \text{otherwise.} \end{cases}$
- The Brier score is $\frac{1}{n}\sum(p_i - x_i)^2$
- (For multiple observations, take average)
- This is a penalty – you are asked to minimize your Brier score, but the only way you can do this is to state your own probabilities.
- There is no benefit to “gaming” by stating something you do not believe

Brier score property

- If I really think $q_1, \ldots, q_n$ but state $p_1, \ldots, p_n$, then

$$\begin{aligned} {}^1\!/_n \textstyle\sum (p_i - x_i)^2 &= {}^1\!/_n \textstyle\sum (p_i^2 - 2p_i x_i + x_i^2) \\ &= {}^1\!/_n \textstyle\sum \left(p_i^2 - 2p_i x_i - q_i^2 + q_i^2 + x_i^2\right) \\ &= {}^1\!/_n \textstyle\sum \left(p_i^2 - 2p_i x_i - q_i^2\right) + {}^1\!/_n \textstyle\sum q_i^2 + 1 \end{aligned}$$

- On average this would be

$$ {}^1\!/_n \sum (p_i - q_i)^2 + {}^1\!/_n \sum q_i^2 + 1 $$

- So to minimize my expected score I should state what I really think!