Evidence-based Decision Making

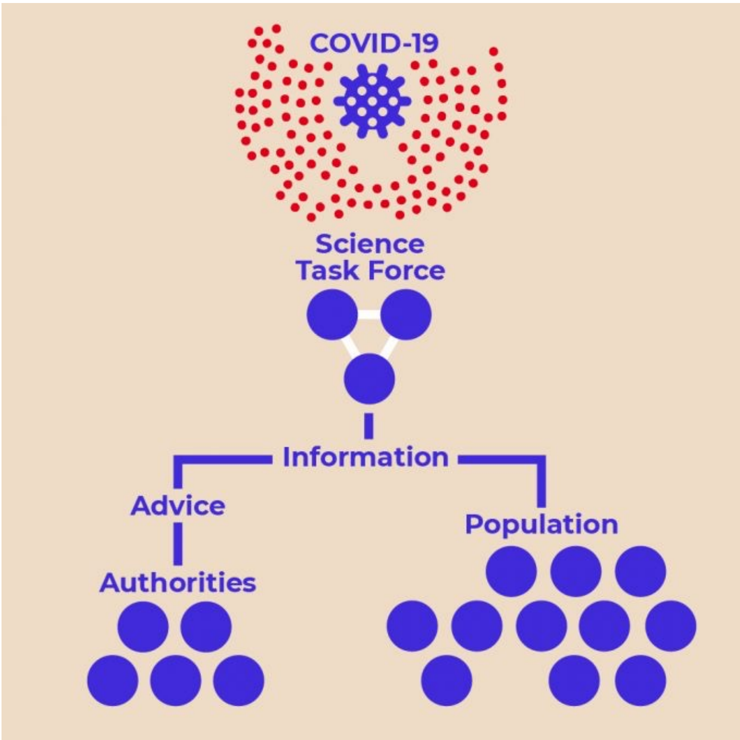
The wisdom of experts

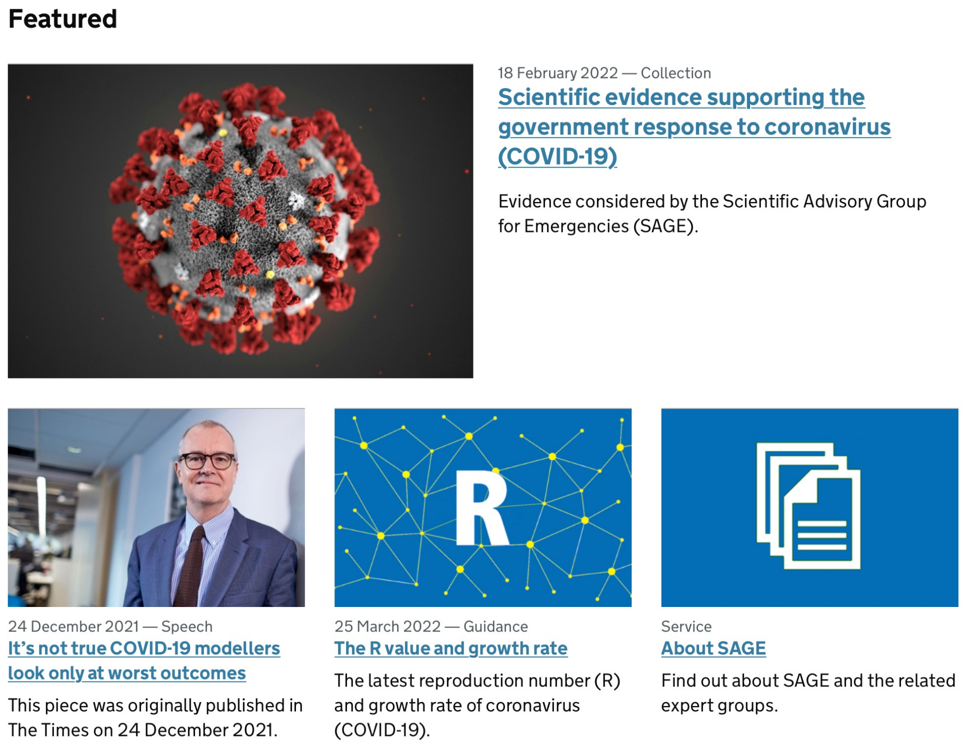
Rui Mata, FS 2023

Version: Feb 15th, 2023

Exercise

What kind of groups are scientific task forces? Can one make recommendations about how experts should interact in these settings?





https://sciencetaskforce.ch/en/home/

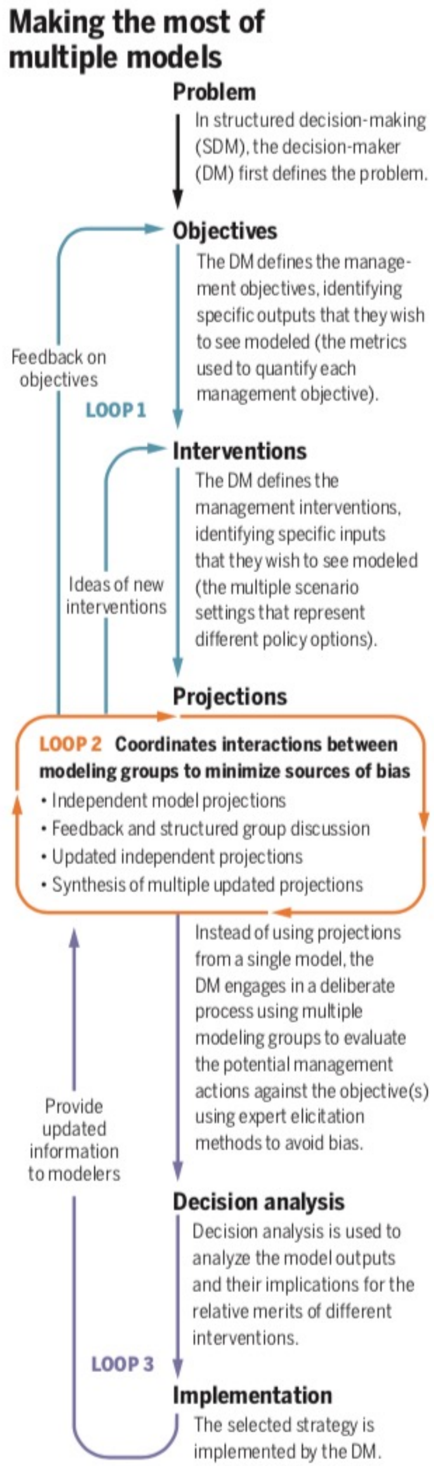
https://www.youtube.com/watch?v=L7uBwyr0sdg

Combining Deliberative and Staticized Groups?

"Disparate predictions during any outbreak can hinder intervention planning and response by policy-makers, who may instead choose to rely on single trusted sources of advice, or on consensus where it appears. (…)

To harness both the creativity of individuals and the insights of groups, variations on the Delphi method (developed by the RAND Corporation in the 1950s and included within the IDEA protocol) and the Nominal Group Technique involve both independent and interactive stages in an iterative elicitation process. The expert judgment literature shows that a failure to manage the elicitation process well can lead to generation of biased information and overconfidence. Expert judgment approaches have been used for elicitation from individual experts in a wide range of relevant settings, such as development of clinical guidelines, and in conservation and ecology."

Shea, K., Runge, M. C., Pannell, D., Probert, W. J. M., Li, S.-L., Tildesley, M., & Ferrari, M. (2020). Harnessing multiple models for outbreak management. *Science*, *368*(6491), 577–579. http://doi.org/10.1126/science.abb9934



Goals

Understanding the performance of groups as a process of statistical aggregation and learn about how to predict when crowds vs. experts will do best.

Learn about how psychology is using the tools of aggregation/consensus to change the way economic and political forecasting is conducted.

Debate possible implications for application to societal issues…

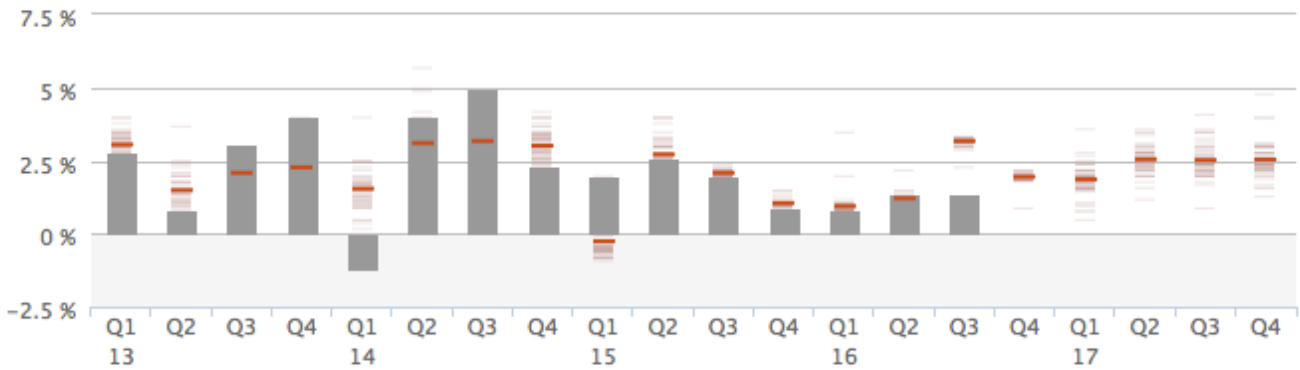
THE WALL STREET JOURNAL.

Economic Forecasting Survey

The Wall Street Journal surveys a group of more than 60 economists on more than 10 major economic indicators on a monthly basis.

GDP (quarterly)

Actual Estimates 7 yr. 5 yr. 3 yr.



Share view:
Embed

GDP (quarterly)

Actual (Q3 2016)	Projected: Q4 2016	Projected: Q1 2017	Projected: Q2 2017
1.4%	1.9% ▲	1.9%	2.5%

http://projects.wsj.com/econforecast/#ind=gdp&r=20

Mannes et al. (2014)

1. simulations to show the relative performance of crowds, best judge, or select crowds as a function of enviroment/judge performance

2. show the relative performance of crowds, best judge, or select crowds in real environments

3. surveys/experiments to evaluate people's intuitions about the performance of staticized groups (crowds, select crowds) vs. best judge

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology, 107*(2), 276–299.

Aggregation of inferences

Mannes et al. (2014) suggest the success of aggregation relative to a best member (expert) or a team of experts depends on the distribution of knowledge (dispersion) and population bias (bracketing)…

Dispersion in expertise: degree to which members differ in ability to estimate the criterion
Bracketing: frequency with which any two judges fall on opposite sides of the criterion

	Low dispersion in expertise	**High dispersion in expertise**
High bracketing	(A) Whole Crowd	(B) Select Crowd
Low bracketing	(C) Select Crowd	(D) Best Member

Figure 1. Four exemplar judgment environments and the strategies expected to perform the best in each.

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology, 107*(2), 276–299.

Aggregation of inferences: Simulations

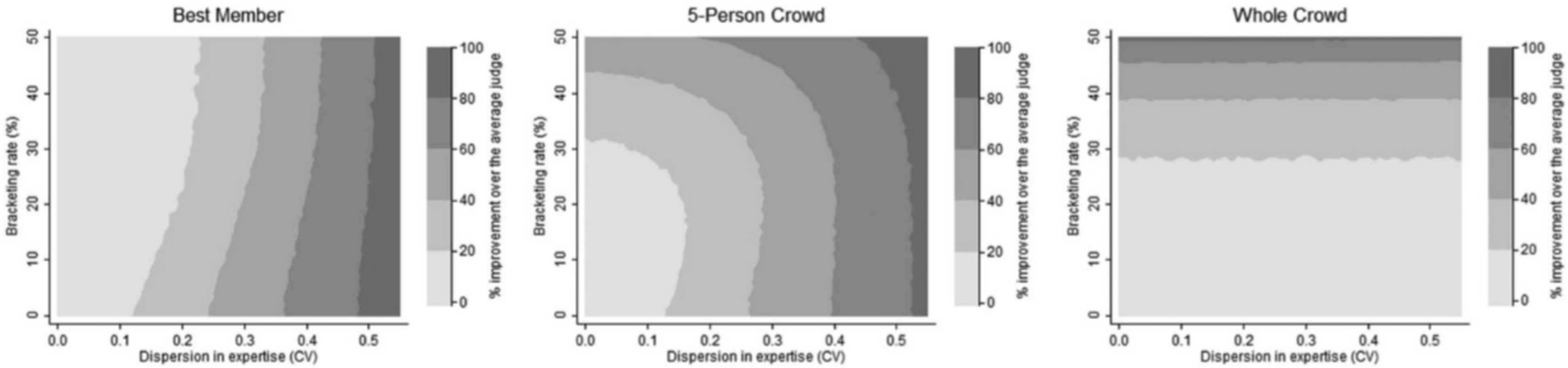


Figure 3. Contour maps of performance across 2,856 simulated judgment environments for three judgment strategies. Five trials of history were used to rank and select judges (N = 50). Darker shades of gray indicate greater percent improvement over the average judge. CV = coefficient of variation.

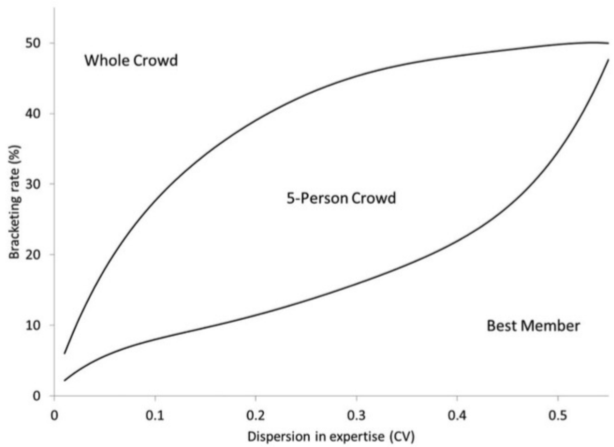


Figure 4. Best-performing strategy for each simulated judgment environment with N = 50 judges ranked and selected based on five periods of history. With less (more) history available to select judges, the curves rotate clockwise (counterclockwise). CV = coefficient of variation.

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology, 107*(2), 276–299.

Aggregation of inferences: Real data

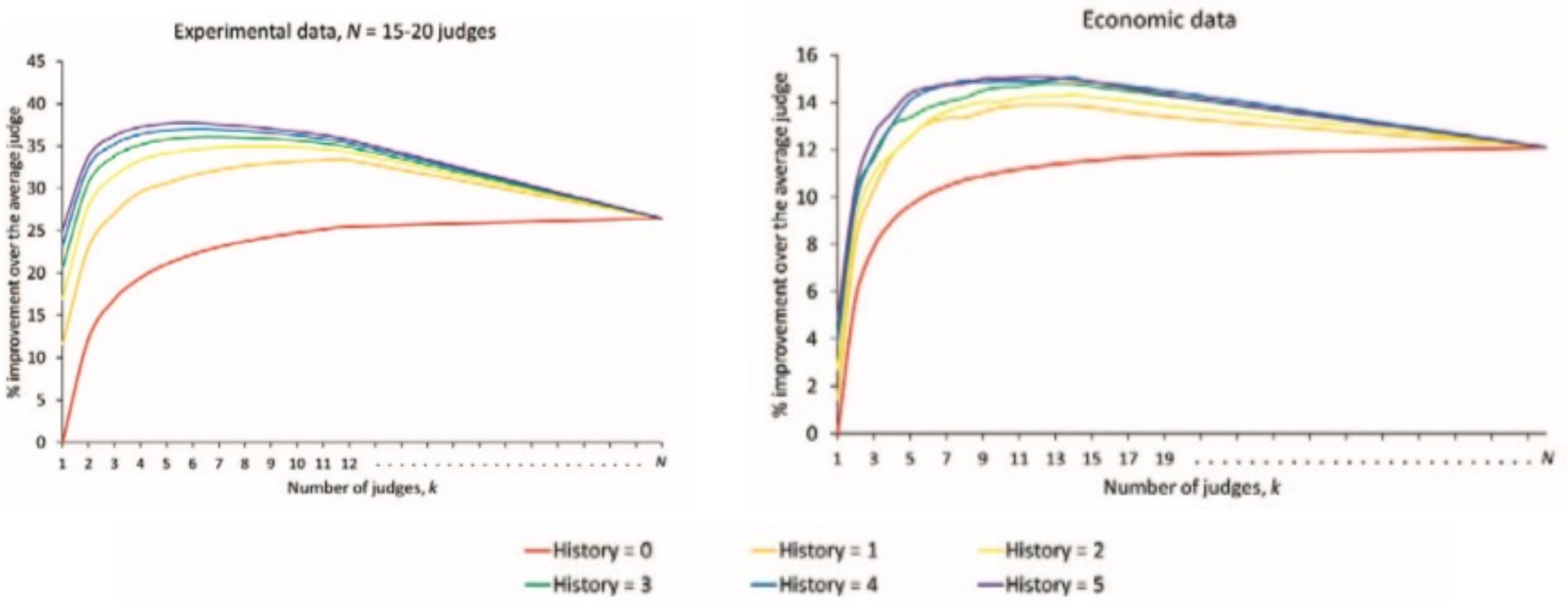


Figure 6. Performance of the best member ($k = 1$), select crowds ($1 < k < N$), and whole crowd ($k = N$) in the experimental and economic data. The experimental data is separated into the 20 sets with $N = 15$–20 judges (top left) and the 20 sets with $N > 20$ judges (top right). Selections based on six levels of history are provided. Performance for omitted values of k (in ellipses) is interpolated.

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology, 107*(2), 276–299.

Aggregation of inferences: Lay intuitions

Table 2
Ratings of Judgment Strategies in Experiment 1

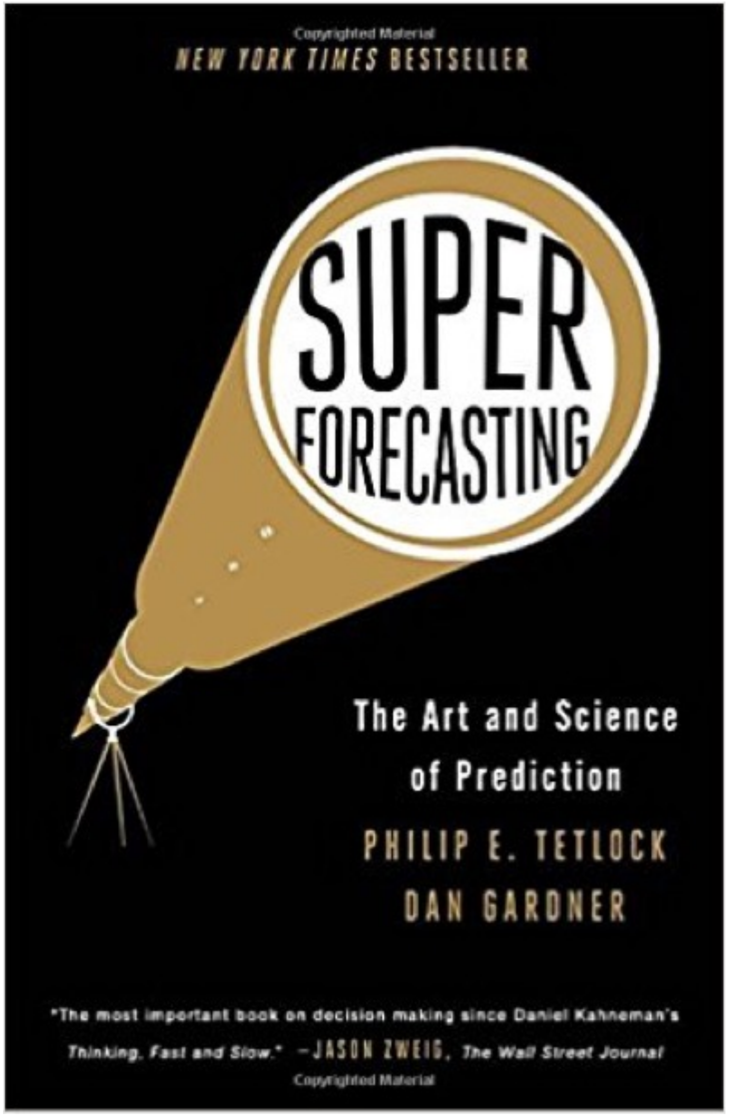
			Difference in means				
Strategy	*M*	*SD*	1	2	3	4	5
1. Random economist	3.24	1.37	—				
2. Average of all economists	4.71	1.22	1.46***	—			
3. Most accurate economist last year	4.60	1.28	1.35***	−0.11	—		
4. Most accurate economist last 5 years	5.04	1.22	1.79***	0.33***	0.44***	—	
5. Average of 5 most accurate economists last year	5.11	1.20	1.86***	0.40***	0.51***	0.07	—

Note. $N = 312$. Mean rating (1 = *not at all accurate* to 7 = *extremely accurate*)
*** $p < .005$ (Bonferroni-adjusted, $\alpha_{FW} = .05$).

People seem to have the intuition that the most accurate expert or a team of experts are about the same…
Possible reasons are beliefs about the (lack of) predictability of judges' future performance rather than beliefs about the power of averaging.

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds.
Journal of Personality and Social Psychology, 107(2), 276–299.

Good Judgment Project



Welcome to

Good Judgment® Open

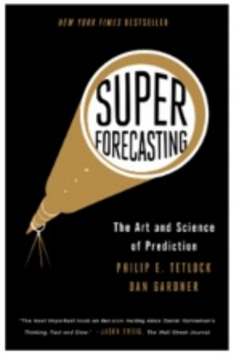
Are you a Superforecaster®?

Join the internet's smartest crowd. Improve your forecasting skills and find out how you stack up.

Forecasting challenge sponsors — including, among others, CNN's *Fareed Zakaria GPS*, *The Economist*, and the University of Pennsylvania's *Mack Institute* — invite you to anticipate the major political, economic, and technological events that will shape 2018.

GPS FAREED ZAKARIA | The Economist | Penn

Be sure to check out all of our active challenges, our featured questions, and our unfiltered list of all open forecasting questions.



About Us

Good Judgment Open is owned and operated by Good Judgment, a forecasting services firm that equips corporate and government decision makers with the benefit of foresight.

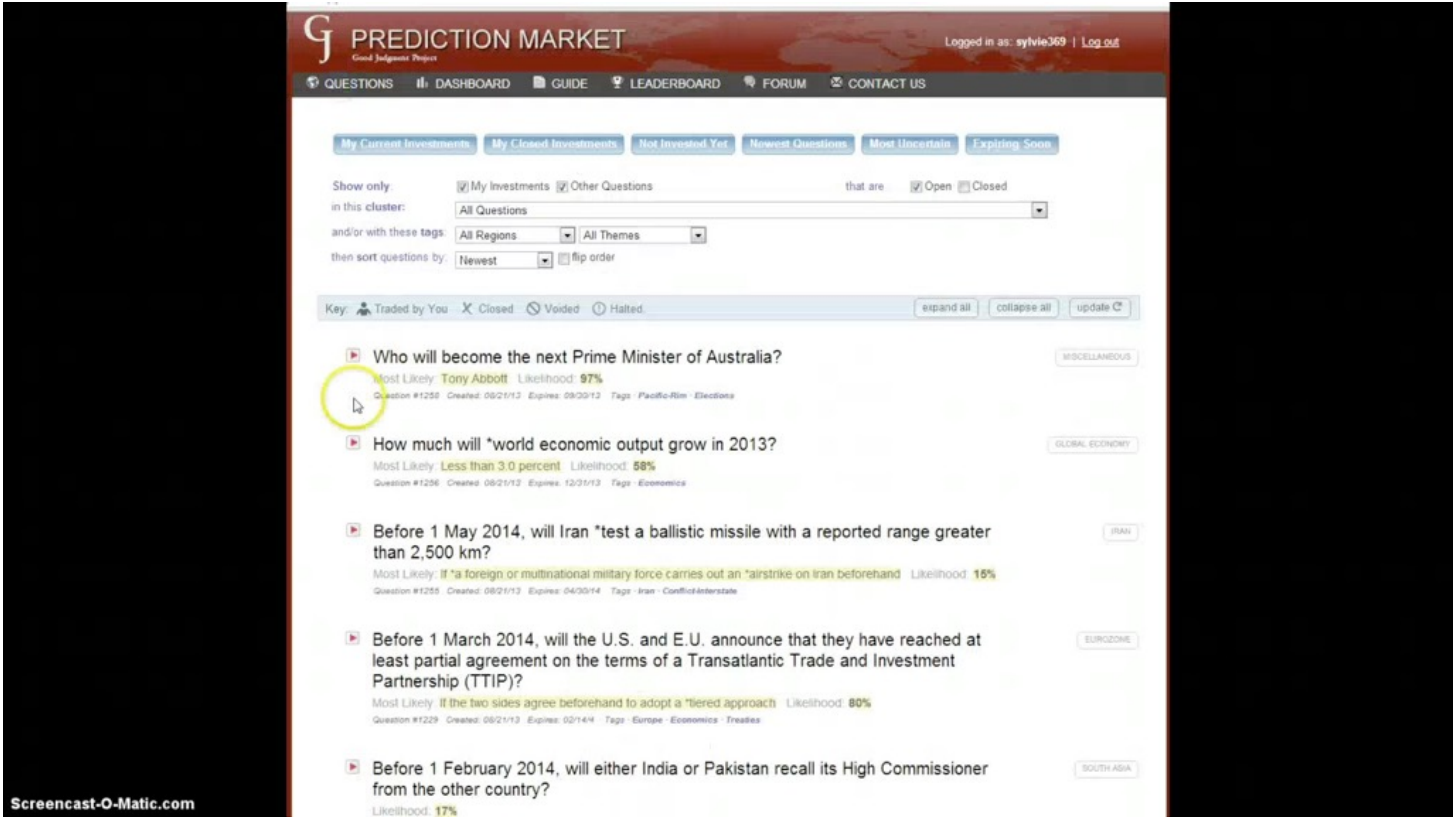
Good Judgment's co-founder, Philip Tetlock, literally wrote the book on state-of-the-art crowd-sourced forecasting. Learn more about Good Judgment and the services it provides at goodjudgment.com.

GOOD JUDGMENT

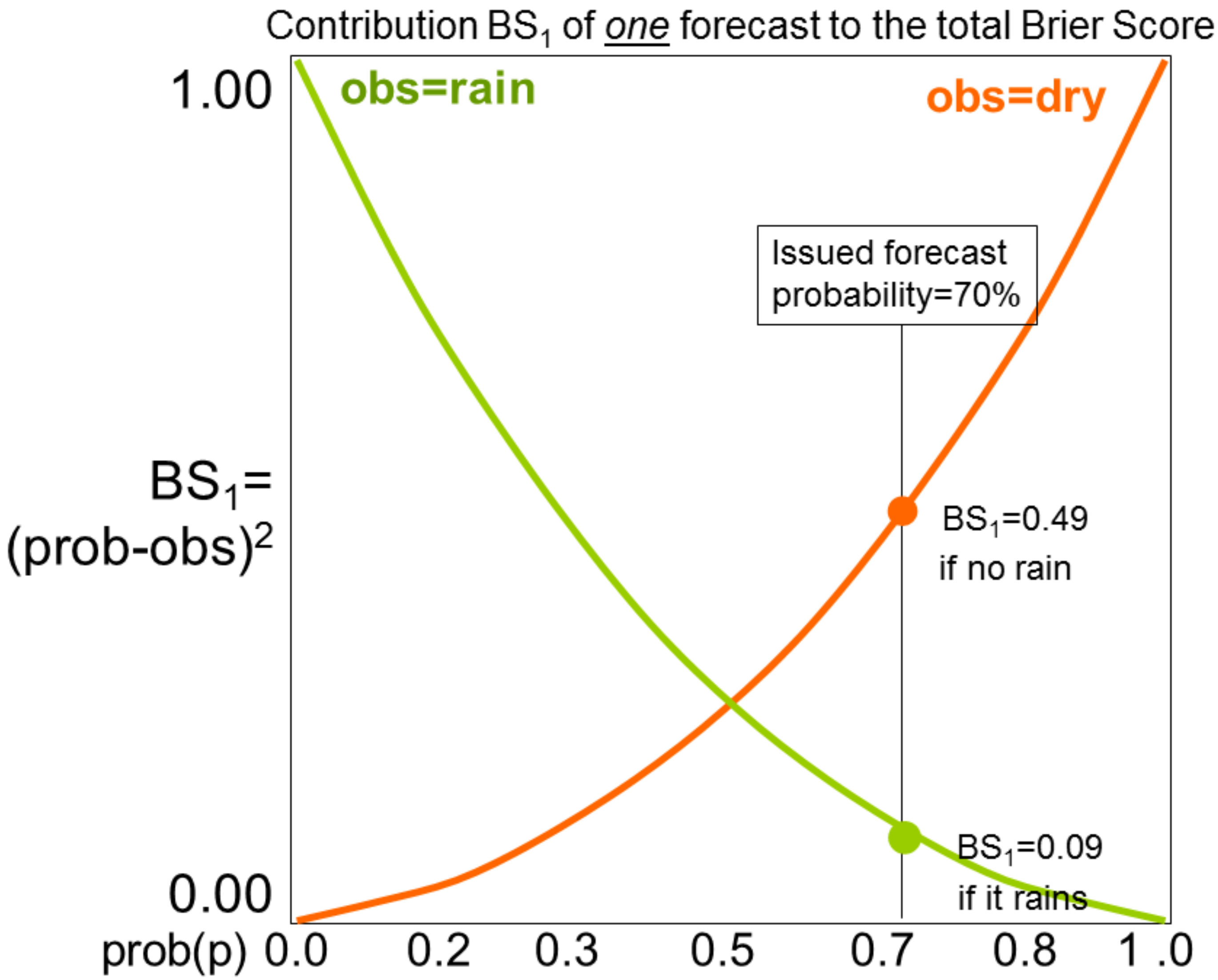
Sign Up

https://goodjudgment.com

Good Judgment Project



Good Judgment Project



https://en.wikipedia.org/wiki/Brier_score

Good Judgment Project

Abstract

Five university-based research groups competed to recruit forecasters, elicit their predictions, and aggregate those predictions to assign the most accurate probabilities to events in a 2-year geopolitical forecasting tournament. Our group tested and found support for three psychological drivers of accuracy: training, teaming, and tracking. Probability training corrected cognitive biases, encouraged forecasters to use reference classes, and provided forecasters with heuristics, such as averaging when multiple estimates were available. Teaming allowed forecasters to share information and discuss the rationales behind their beliefs. Tracking placed the highest performers (top 2% from Year 1) in elite teams that worked together. Results showed that probability training, team collaboration, and tracking improved both calibration and resolution. Forecasting is often viewed as a statistical problem, but forecasts can be improved with behavioral interventions. Training, teaming, and tracking are psychological interventions that dramatically increased the accuracy of forecasts. Statistical algorithms (reported elsewhere) improved the accuracy of the aggregation. Putting both statistics and psychology to work produced the best forecasts 2 years in a row.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., et al. (2014). Psychological Strategies for Winning a Geopolitical Forecasting Tournament. *Psychological Science*, *25*(5), 1106–1115. http://doi.org/10.1177/0956797614524255

Good Judgment Project

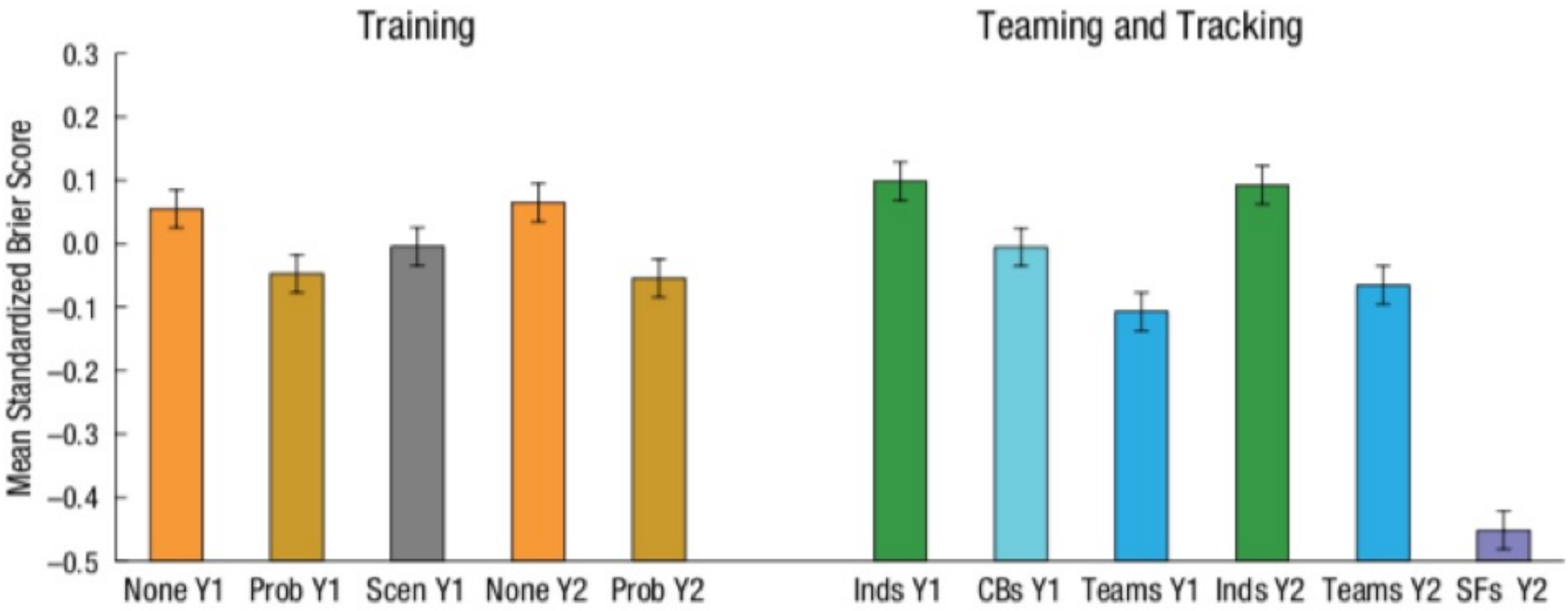


Fig. 1. Effects of training, teaming, and tracking on average Brier scores in Year 1 (Y1) and Year (Y2). The bars at the left show results for the no-training ("None"), probability-training ("Prob"), and scenario-training ("Scen") conditions; the bars at the right show results for independent forecasters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

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., et al. (2014). Psychological Strategies for Winning a Geopolitical Forecasting Tournament. *Psychological Science*, *25*(5), 1106–1115. http://doi.org/10.1177/0956797614524255

Good Judgment Project: Superforecasters

Table 3. Correlates With Measures With Accuracy

Measure	Correlation	$t(1774)$	p
Raven's Advanced Progressive Matrices	−.18	−7.70	<.001
Shipley–2 Abstraction Test	−.22	−9.49	<.001
Shipley–2 Vocabulary	−.09	−3.80	<.001
CRT	−.16	−6.82	<.001
Extended CRT	−.23	−9.95	<.001
Numeracy	−.16	−6.82	<.001
Political knowledge (Year 1)	−.12	−5.09	<.001
Political knowledge (Year 2)	−.18	−7.70	<.001
Political knowledge (Year 3)	−.14	−5.95	<.001
Motivate—Be at the top	−.11	−4.66	<.001
Need for cognition	−.07	−2.95	<.002
Active open-mindedness	−.12	−5.09	<.001
Average number of articles checked	−.18	−7.70	<.001
Average number of articles shared	−.20	−8.53	<.001
Average number of comments with questions	−.18	−7.68	<.001
Average number of replies to questions	−.18	−7.70	<.001

Note: CRT = Cognitive Reflection Test.

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., et al. (2015). Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions. ***Perspectives on Psychological Science, 10***(3), 267–281. http://doi.org/10.1177/1745691615577794

Good Judgment Project: Superforecasters

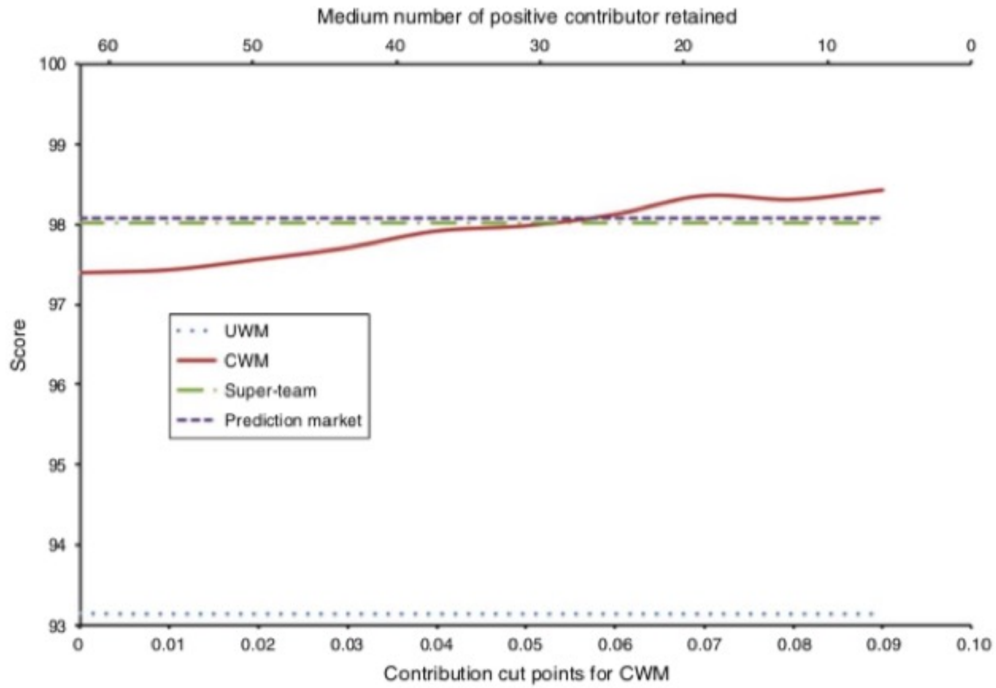
Abstract

Across a wide range of tasks, research has shown that people make poor probabilistic predictions of future events. Recently, the U.S. Intelligence Community sponsored a series of forecasting tournaments designed to explore the best strategies for generating accurate subjective probability estimates of geopolitical events. In this article, we describe the winning strategy: culling off top performers each year and assigning them into elite teams of *superforecasters*. Defying expectations of regression toward the mean 2 years in a row, superforecasters maintained high accuracy across hundreds of questions and a wide array of topics. We find support for four mutually reinforcing explanations of superforecaster performance: (a) cognitive abilities and styles, (b) task-specific skills, (c) motivation and commitment, and (d) enriched environments. These findings suggest that superforecasters are partly discovered and partly created—and that the high-performance incentives of tournaments highlight aspects of human judgment that would not come to light in laboratory paradigms focused on typical performance.

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., et al. (2015). Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions. ***Perspectives on Psychological Science, 10***(3), 267–281. http://doi.org/10.1177/1745691615577794

A Better Crystal Ball: Good judgment project & select-crowds

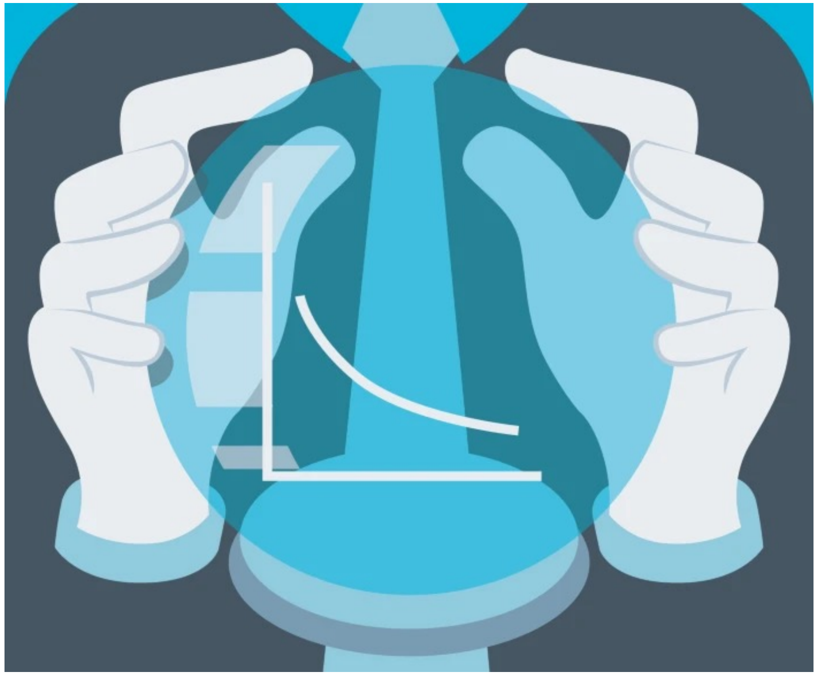
Figure 2 (Color online) CWM Beats Super-Teams and Prediction Market by Increasing the Threshold for Selecting Positive Contributors in Period 1



The Contribution-Weighted Model (CWM) is a version of the select-crowd strategy – it measures the relative contribution of each judge to the group and weights each judge according to their past performance; interestingly, results suggest it is more important to identify the best judges than to weigh each one appropriately.

Chen, E., Budescu, D. V., Lakshmikanth, S. K., Mellers, B. A., & Tetlock, P. E. (2016). Validating the Contribution-Weighted Model: Robustness and Cost-Benefit Analyses. *Decision Analysis, 13*(2), 128–152. http://doi.org/10.1287/deca.2016.0329

A Better Crystal Ball: Integration of approaches



Scenario planning

e.g., planners create critical uncertainties and, taking the extreme values, constructing possible future worlds (2 x 2 matrix)

Probabilistic forecasting

e.g., forecaster use logic and calculation to describe the behavior of a system and predict (assign a probability) to a future state

Scoblic, P., & Tetlock, P.E. (2020). A Better Crystal Ball. *Foreign Affairs*, Nov/Dec, 99, 6.

A Better Crystal Ball: The inner crowd…

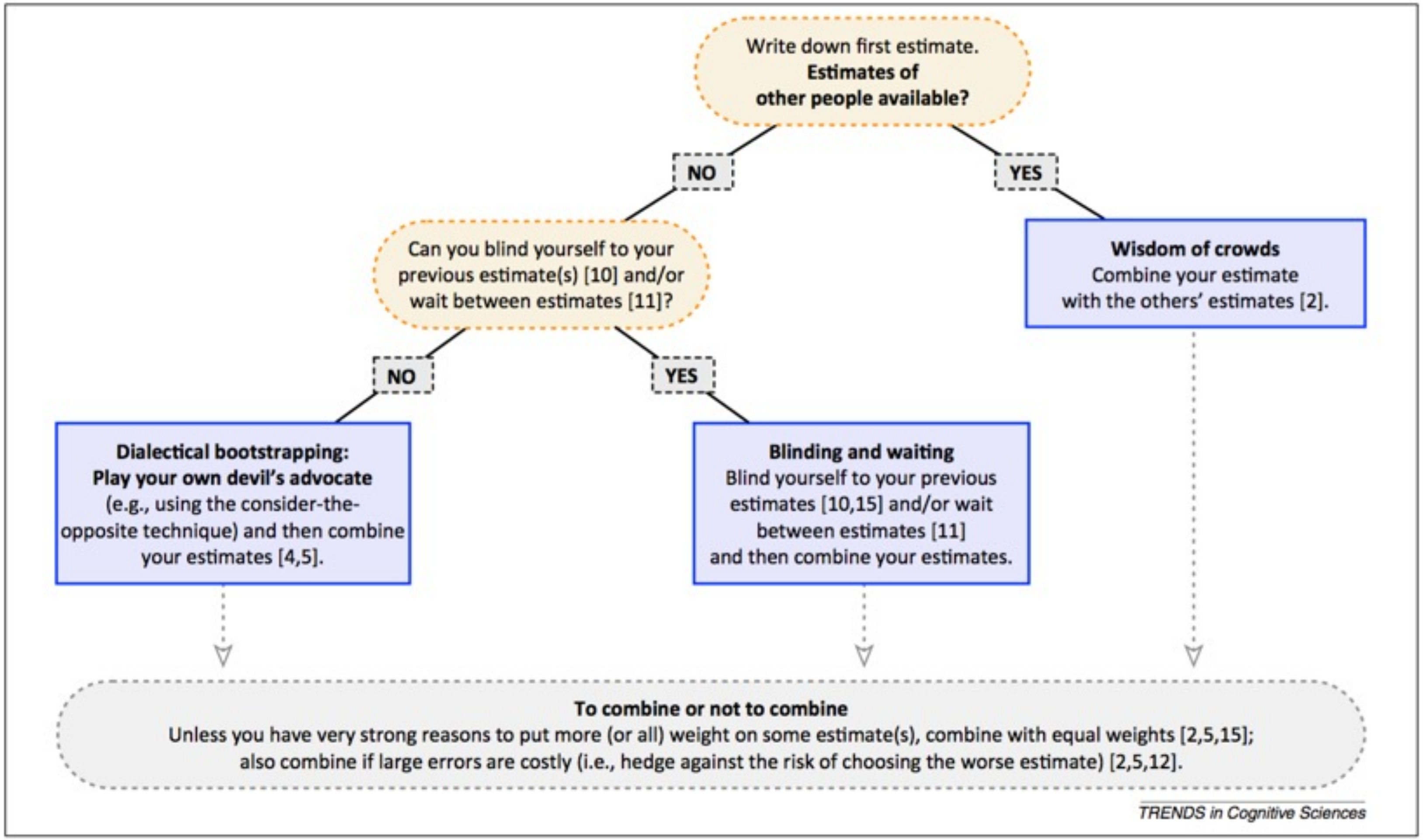


Figure 1. Decision tree for deciding when and how to use the inner crowd.

Herzog, S. M., & Hertwig, R. (2014). Harnessing the wisdom of the inner crowd. *Trends in Cognitive Sciences, 18*(10), 504–506.

Implications…

DiversiTy

@ University of Basel 2000-2020

https://cdsbasel.github.io/Diversity_hackathon/

Demographically diverse crowds are not much wiser than homogeneous crowds

Table 2. Very homogeneous and diverse groups

Task	G1*	G2	G3	G4	G5	G6	G7
Predict percentage of votes eight presidential candidates would receive in two state primaries	Random	White men, did not complete college	White women, completed college	Religious white Republican	Nonreligious white Democrats	Liberal women under 40	Liberal nonwhites
Guess what percentage of Americans support each of six political statements	Random	White men, did not complete college	White women, completed college	Religious white conservative	Nonreligious white liberals	Liberal women under 40	Liberal nonwhites
Predict what percentage of votes Clinton and Trump would each win in 10 states in 2016 presidential election	Random	White men, did not complete college	White women, completed college	Religious white Republican	Nonreligious white Democrats	Liberal women over 40	Liberal nonwhites
Guess the popularity rating that 24 diverse books received in a previous study	Random	Men over 40	Men under 30	Women over 40	Women under 30	Ethnic minority women	White men

G1* is always the diverse crowd. All groups except two were simulated from pools of at least 30 people. G2 for the book task (men over 40) was simulated from a pool of 22 men, and G6 (ethnic minority women) was sampled from a pool of 29 due to limited representation of those groups in the larger sample.

De Oliveira, S., & Nisbett, R. E. (2018). Demographically diverse crowds are typically not much wiser than homogeneous crowds. *Proceedings of the National Academy of Sciences, 115*(9), 2066-2071. https://doi.org/10.1073/pnas.1717632115

Demographically diverse crowds are not much wiser than homogeneous crowds

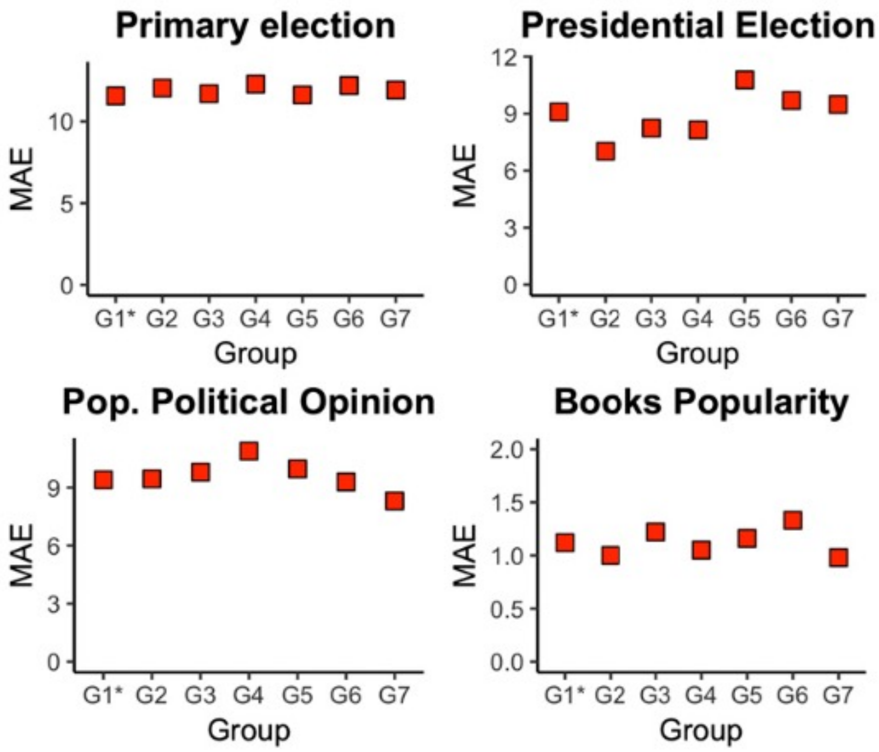


Fig. 2. Very homogeneous vs. very diverse groups across four tasks. G1 always represents the diverse crowds. G2 to G7 represent the homogeneous groups as described in Table 2. For example, for the primary election, presidential election, and popular political opinion tasks, G2 refers to white men who did not complete college. For the book-rating task, G2 refers to men over 40 y old. In all graphs, the *y* axis indicates error. Lower values mean higher accuracy on the task.

De Oliveira, S., & Nisbett, R. E. (2018). Demographically diverse crowds are typically not much wiser than homogeneous crowds. *Proceedings of the National Academy of Sciences, 115*(9), 2066-2071. https://doi.org/10.1073/pnas.1717632115

Implications…

“Despite advocates’ insistence that women on boards enhance corporate performance and that diversity of task groups enhances their performance, research findings are mixed, and repeated meta-analyses have yielded average correlational findings that are null or extremely small. Therefore, social scientists should (a) conduct research to identify the conditions under which the effects of diversity are positive or negative and (b) foster understanding of the social justice gains that can follow from diversity. Unfortunately, promulgation of false generalizations about empirical findings can impede progress in both of these directions. Rather than ignoring or furthering distortions of scientific knowledge to fit advocacy goals, scientists should serve as honest brokers who communicate consensus scientific findings to advocates and policy makers in an effort to encourage exploration of evidence-based policy options.”

Eagly, A. H. (2016). When passionate advocates meet research on diversity, does the honest broker stand a chance. Journal of Social Issues, 72(1), 199-222. https://doi.org/10.1111/josi.12163

Summary

Staticized groups can work well. Understanding the performance of groups as a process of statistical aggregation involving different factors - dispersion and bracketing - helps predict when select crowds (or other types of aggregation) will do best.

Aggregating preferences over a whole crowd works best when there is low dispersion of knowledge and high bracketing. Trusting a single expert makes sense if he/she has all the knowledge!

Often, teams of experts seem to provide a good balance by capitalising on dispersion and bracketing. Lay people are not fully aware of the power of aggregation and of select crowds…

Beware of drawing of implications for diversity management: the literature is not yet mature but many mixed findings concerning diversity for performance… We should argue for diversity based on ethical, not performance grounds!

COMMENTARY Open Access

Reimagining peer review as an expert elicitation process

Alexandru Marcoci[1*], Ans Vercammen[2,6], Martin Bush[3], Daniel G. Hamilton[3], Anca Hanea[3,4], Victoria Hemming[5], Bonnie C. Wintle[3], Mark Burgman[6] and Fiona Fidler[3]

Abstract

Journal peer review regulates the flow of ideas through an academic discipline and thus has the power to shape what a research community knows, actively investigates, and recommends to policymakers and the wider public. We might assume that editors can identify the 'best' experts and rely on them for peer review. But decades of research on both expert decision-making and peer review suggests they cannot. In the absence of a clear criterion for demarcating reliable, insightful, and accurate expert assessors of research quality, the best safeguard against unwanted biases and uneven power distributions is to introduce greater transparency and structure into the process. This paper argues that peer review would therefore benefit from applying a series of evidence-based recommendations from the empirical literature on structured expert elicitation. We highlight individual and group characteristics that contribute to higher quality judgements, and elements of elicitation protocols that reduce bias, promote constructive discussion, and enable opinions to be objectively and transparently aggregated.

Keywords: Peer review, Expert elicitation, Wisdom of the crowd, Anonymity, DELPHI

Marcoci, A., Vercammen, A., Bush, M., Hamilton, D. G., Hanea, A., Hemming, V., Wintle, B. C., Burgman, M., & Fidler, F. (2022). Reimagining peer review as an expert elicitation process. BMC Research Notes, 15(1), 127. https://doi.org/10.1186/s13104-022-06016-0

Fig. 1 The IDEA protocol for structured expert judgement elicitation (adapted from [20])

Marcoci, A., Vercammen, A., Bush, M., Hamilton, D. G., Hanea, A., Hemming, V., Wintle, B. C., Burgman, M., & Fidler, F. (2022). Reimagining peer review as an expert elicitation process. BMC Research Notes, 15(1), 127. https://doi.org/10.1186/s13104-022-06016-0