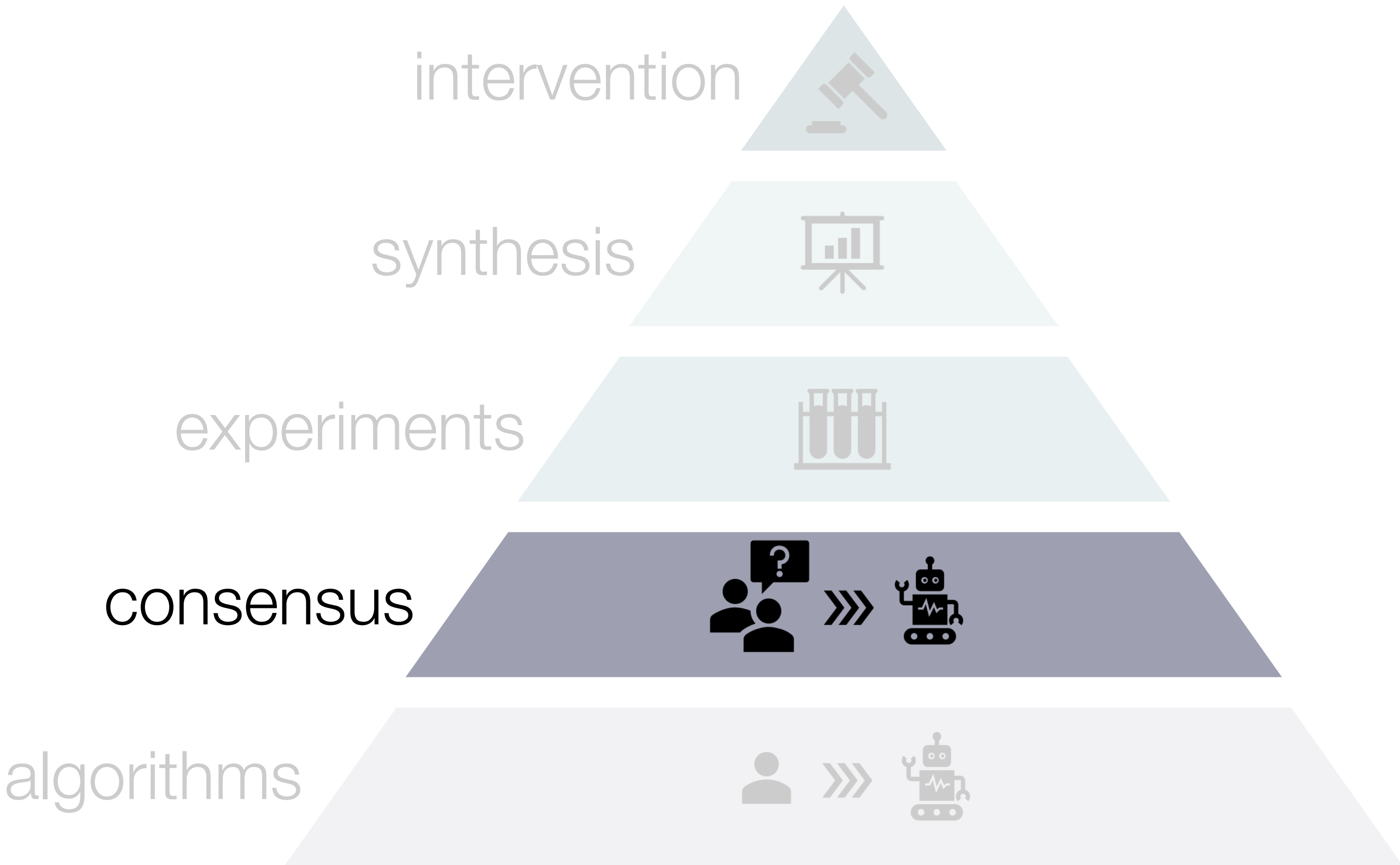
Evidence-based Decision Making
Consensus: The wisdom of (expert) crowds

Loreen Tisdall, FS 2026

Version: March 23, 2026

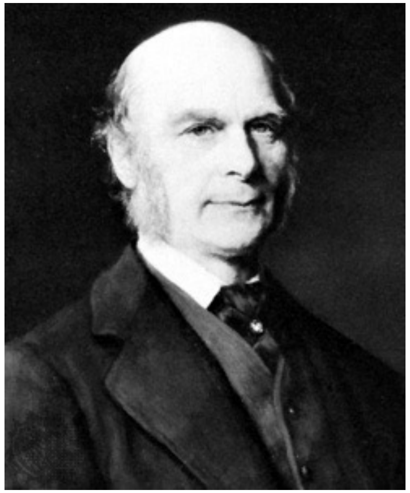
Climbing the pyramid of evidence



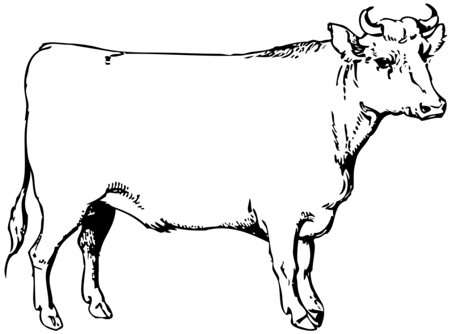
Goals for today

- Understand the performance of groups as a process of statistical aggregation
- Learn about when crowds vs. experts vs. select crowds will do best
- Learn about how psychology is using the tools of aggregation/consensus to change the way economic and political forecasting is conducted

When groups work: Wisdom of the crowd!



Francis Galton, 1822-1911



True weight = 1198 pounds

“This result is, I think, more credible to the trustworthiness of a democratic judgment than might have been expected.”

Statisticized groups can be powerful!

Distribution of the estimates of the dressed weight of a particular living ox, made by 787 different persons.

Degrees of the length of Array 0°—100°	Estimates in lbs.	Centiles: Observed deviates from 1207 lbs.	Centiles: Normal p.e =37	Excess of Observed over Normal
5	1074	−133	−90	+43
10	1109	−98	−70	+28
15	1126	−81	−57	+24
20	1148	−59	−46	+13
q_1 25	1162	−45	−37	+8
30	1174	−33	−29	+4
35	1181	−26	−21	+5
40	1188	−19	−14	+5
45	1197	−10	−7	+3
m 50	1207	0	0	0
55	1214	+7	+7	0
60	1219	+12	+14	−2
65	1225	+18	+21	−3
70	1230	+23	+29	−6
q_3 75	1236	+29	+37	−8
80	1243	+36	+46	−10
85	1254	+47	+57	−10
90	1267	+52	+70	−18
95	1293	+86	+90	−4

q_1, q_3, the first and third quartiles, stand at 25° and 75° respectively.
m, the median or middlemost value, stands at 50°.
The dressed weight proved to be 1198 lbs.

Galton, F. (1907). Vox Populi. *Nature,* 75, 450-451.

Why groups work

A BIOLOGIST, A CHEMIST, AND
A STATISTICIAN ARE OUT HUNTING.
THE BIOLOGIST SHOOTS AT A DEER
AND MISSES 5FT TO THE LEFT, THE
CHEMIST TAKES A SHOT AND MISSES
5FT TO THE RIGHT, THE STATISTICIAN
YELLS "WE GOT 'EM!"

Not just your average kind of joke ;)

Your turn!

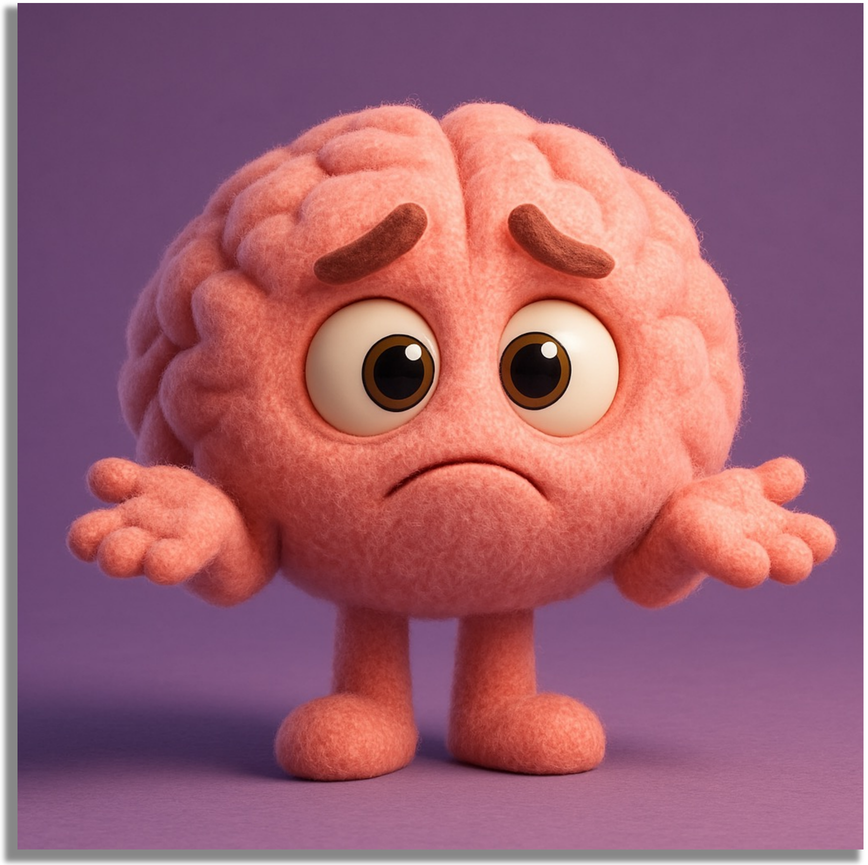


Image created with AI (ChatGPT 4o), March 29, 2025

In which areas of (your) life do you come across consensus-based judgments?

Discuss with your neighbour(s)

~2 minutes

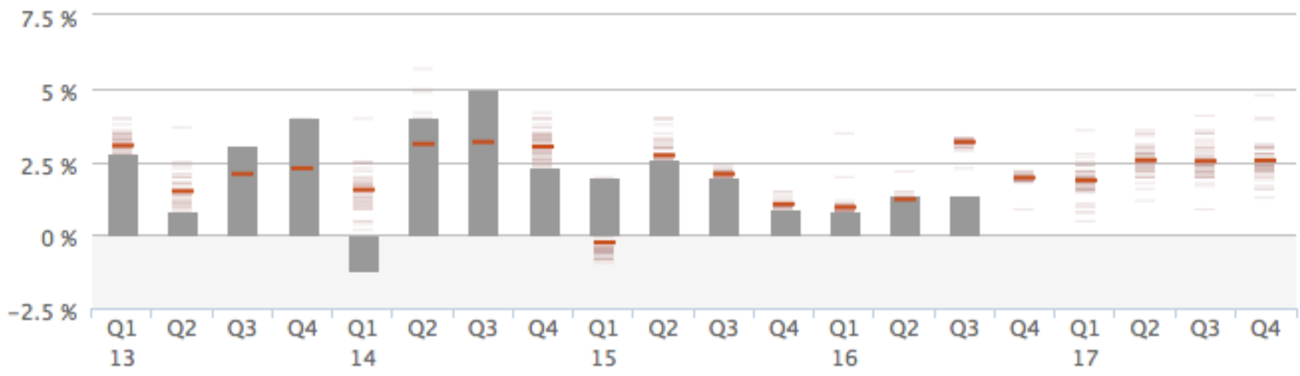
THE WALL STREET JOURNAL.

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%

https://www.wsj.com/economy/economic-forecasting-survey-archive-11617814998

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.

→ Whose opinion **should** people follow if they desire to maximize their accuracy, and whose **do** they follow when making these decisions?

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%

https://www.wsj.com/economy/economic-forecasting-survey-archive-11617814998

Key reading:
The wisdom of best judge, crowds, select crowds

Image created with AI (ChatGPT 4o), March 29, 2025

Journal of Personality and Social Psychology
2014, Vol. 107, No. 2, 276–299

0022-3514/14/$12.00 http://dx.doi.org/10.1037/a0036677

The Wisdom of Select Crowds

Albert E. Mannes
Philadelphia, Pennsylvania

Jack B. Soll and Richard P. Larrick
Duke University

Social psychologists have long recognized the power of *statisticized* groups. When individual judgments about some fact (e.g., the unemployment rate for next quarter) are averaged together, the average opinion is typically more accurate than most of the individual estimates, a pattern often referred to as the *wisdom of crowds*. The accuracy of averaging also often exceeds that of the individual perceived as most knowledgeable in the group. However, neither averaging nor relying on a single judge is a robust strategy; each performs well in some settings and poorly in others. As an alternative, we introduce the *select-crowd* strategy, which ranks judges based on a cue to ability (e.g., the accuracy of several recent judgments) and averages the opinions of the top judges, such as the top 5. Through both simulation and an analysis of 90 archival data sets, we show that select crowds of 5 knowledgeable judges yield very accurate judgments across a wide range of possible settings—the strategy is both accurate and robust. Following this, we examine how people prefer to use information from a crowd. Previous research suggests that people are distrustful of crowds and of mechanical processes such as averaging. We show in 3 experiments that, as expected, people are drawn to experts and dislike crowd averages—but, critically, they view the select-crowd strategy favorably and are willing to use it. The select-crowd strategy is thus accurate, robust, and appealing as a mechanism for helping individuals tap collective wisdom.

Keywords: judgment, decision making, aggregation, expertise, groups

Each year, *The Wall Street Journal* conducts a forecasting competition for economists. There are typically around 50 participants representing some of the nation's most elite academic, government, and business institutions. The task is to predict a host of economic variables over the coming year, such as the rates of unemployment and economic growth in the United States. A feature story is later published that celebrates the foresight of the economist who prognosticated the most accurately. Of course, many important choices are influenced by economic forecasts, from companies' hiring decisions to the Federal Reserve Board's position on interest rates. This raises two interesting questions for social scientists—whose opinion do people follow when making these decisions, and whose should they follow if they desire to maximize their accuracy?

We begin this article with the prescriptive question inspired by *The Wall Street Journal*'s contest: How should someone confronted with a set of diverse opinions use them in order to make the best judgment possible? Past social psychological research offers two solutions to this problem. One strategy is to seek out the most knowledgeable person in a group and rely on his or her judgment. This is commonly referred to as a *best-member* strategy (Yetton & Bottger, 1982), and considerable research has focused on ways to improve groups' ability to identify and leverage their expertise (e.g., Bonner, 2004; Hackman, 1987; Henry, 1995; Libby, Trotman, & Zimmer, 1987; Steiner, 1972). Socrates himself was perhaps the strategy's first advocate:

> And for this reason, as I imagine,—because a good decision is based on knowledge and not on numbers? . . . Must we not then first of all ask whether there is any one of us who has knowledge of that about which we are deliberating? If there is, let us take his advice, though he be one only, and not mind the rest; if there is not, let us seek further counsel. (Plato, 2005, p. 46)

This strategy can easily be extended beyond the group context to any situation in which a decision maker is confronted with a crowd of opinions. To judge well, the decision maker should identify the crowd's most qualified member and defer to his or her opinion.

An increasingly popular alternative strategy is to leverage collective knowledge by relying on the *wisdom of crowds* (Surowiecki, 2004). This phenomenon was famously demonstrated by Francis Galton (1907), who reported that the dressed weight of an ox on display at the local fair was only one pound more than the mean estimate of nearly 800 spectators. Since then, exploring the benefits of *statisticized groups* has been a mainstay of social psychological research (e.g., Davis, 1996; Hastie, 1986; Hinsz, 1999; Hogarth, 1978; Lorge, Fox, Davitz, & Brenner, 1958; Stroop, 1932; Wallsten, Budescu, Erev, & Diederich, 1997; Yaniv, 2004). Numerous studies have demonstrated that simple rules for aggregating judgments (such as the median or mean for numerical

Albert E. Mannes, Philadelphia, Pennsylvania; Jack B. Soll and Richard P. Larrick, Fuqua School of Business, Duke University.

This research benefited greatly from the input and feedback we received at numerous seminars and conference presentations. In particular, we wish to acknowledge David Budescu, Robin Hogarth, Cade Massey, Barbara Mellers, Philip Tetlock, and George Wu for their insightful contributions. We also thank Don Moore and his colleagues for sharing data from their studies with us.

Correspondence concerning this article should be addressed to Jack B. Soll, Fuqua School of Business, Duke University, Durham, NC 27708. E-mail: jsoll@duke.edu

***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.

The wisdom of best judge, crowds, select crowds

In this paper, Mannes and colleagues:

- **use simulations** to show the relative performance of crowds, best judge, or select crowds as a function of environment/judge performance
- **show the relative performance** of crowds, best judge, or select crowds **in real environments**
- use **surveys/experiments to evaluate** people's intuitions about the performance of statisticized 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

Expectation (hypothesis): success of aggregation relative to a best judge (expert) or a team of experts (select crowd) depends on the distribution of knowledge (dispersion) and population bias (bracketing)

- **Dispersion in expertise:** degree to which members differ in their ability to estimate the criterion accurately, regardless of the level of expertise (e.g., zero dispersion could be all novices or all experts)
- **Bracketing:** frequency with which any two judges fall on opposite (either) sides of the criterion (correlated / biased error)

A super simplified example to give you an intuition (truth/criterion = 600)

(NOTE: this is not how you actually calculate bracketing rate because we are not considering all possible pairings of judges!!!)

Judge 1	Judge 2
400	800
200	100
550	980
700	900
800	300
599	700
50	1000
550	650
500	700
400	500

- High bracketing → if you pick random pairs of judges and the criterion is frequently between the judges' estimates → good sign for the crowd's diversity of thought
- Low bracketing → if you pick random pairs of judges and the criterion is frequently to one side of the judges' estimates → all guesses may be biased in one direction

***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

Expectation (hypothesis): success of aggregation relative to a best judge (expert) or a team of experts (select crowd) depends on the distribution of knowledge (dispersion) and population bias (bracketing)

- **Dispersion in expertise:** degree to which members differ in their ability to estimate the criterion accurately, regardless of the level of expertise (e.g., zero dispersion could be all novices or all experts)
- **Bracketing:** frequency with which any two judges fall on opposite (either) sides of the criterion (correlated / biased error)

	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.

→ Do select crowds provide a **robust** strategy?

***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 (discrete)

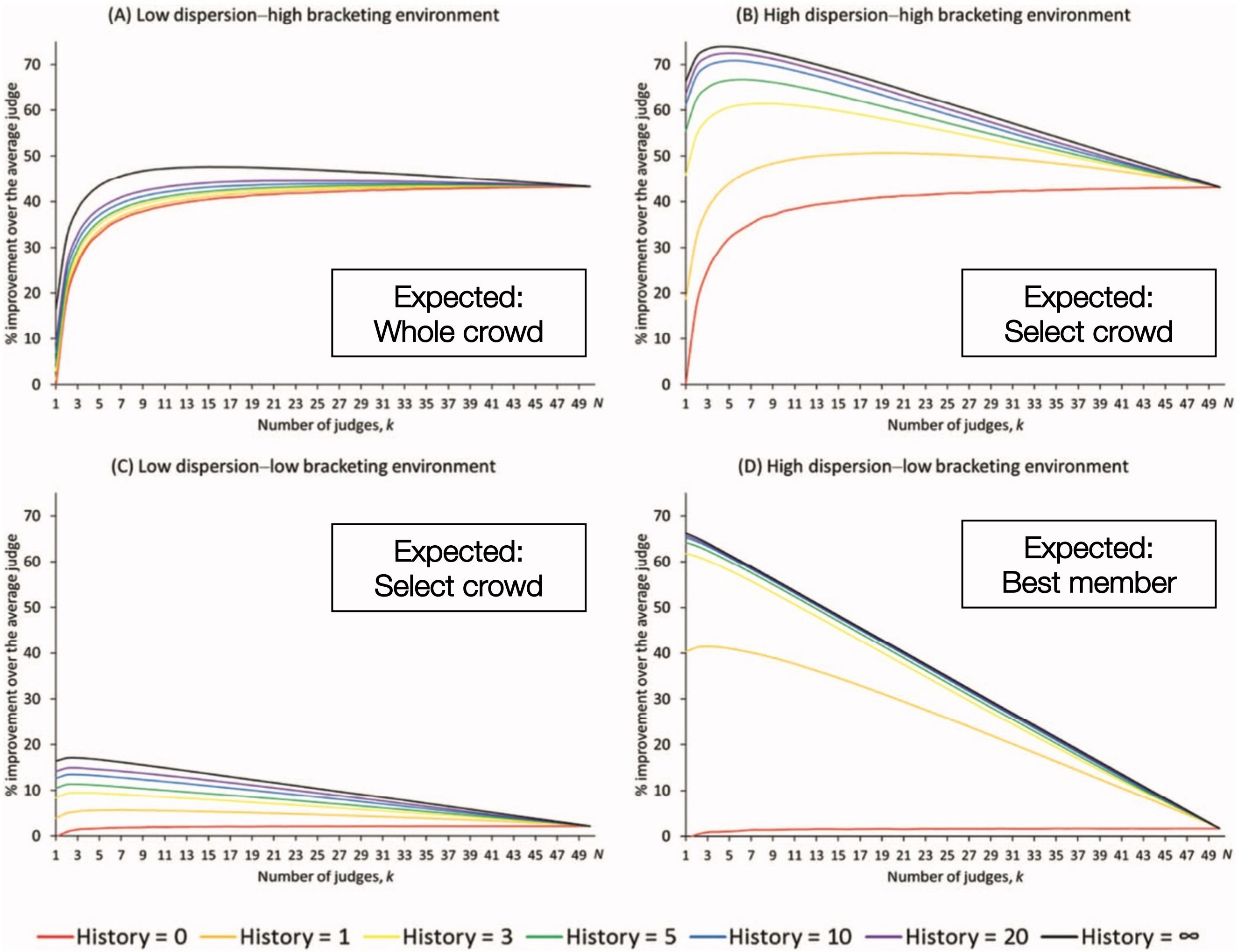
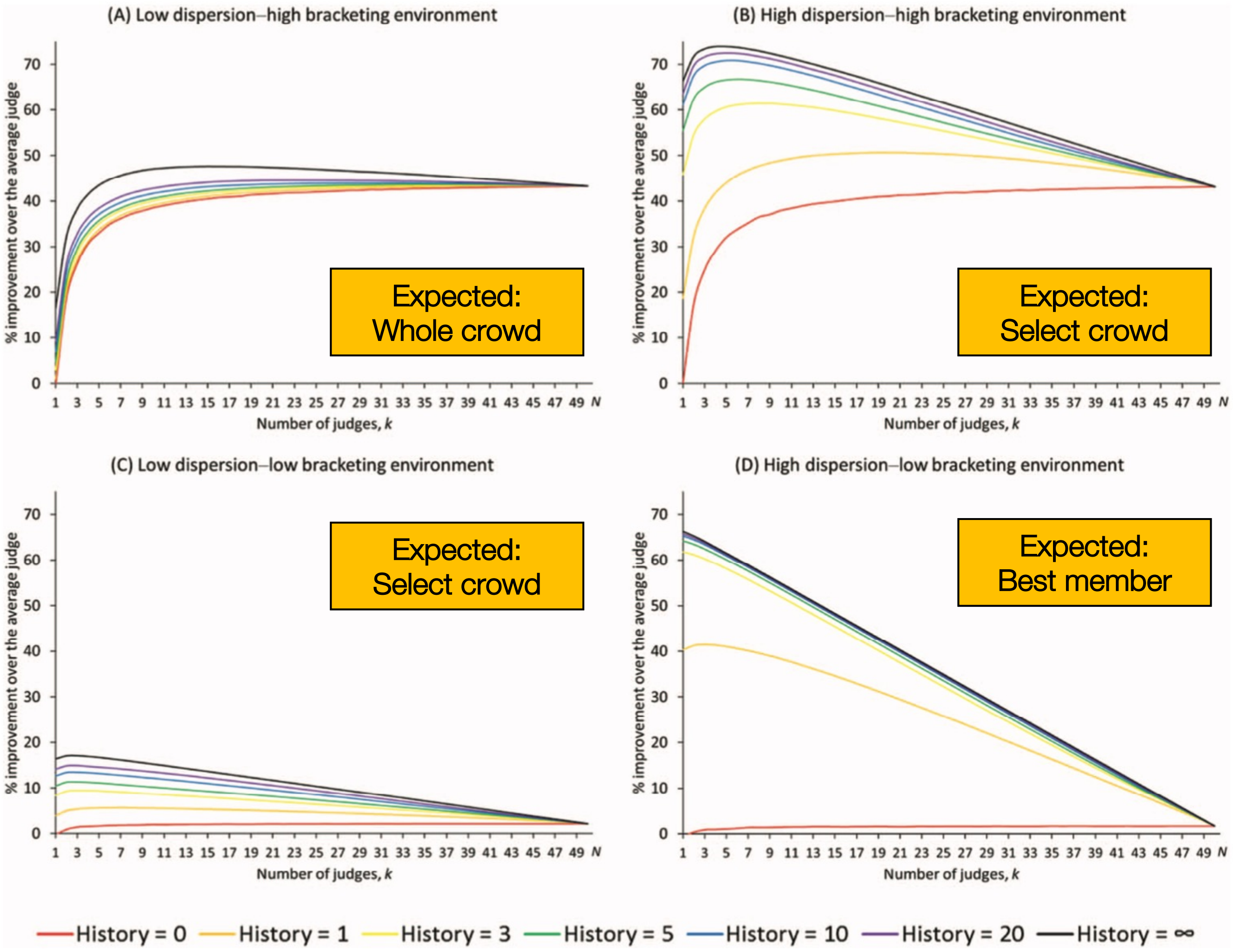


Figure 2. Performance of judgment strategies for a simulated crowd of 50 judges. The performance of the best member is indicated at k 1, of the whole crowd at k N, and of select crowds at 1 k N. Curves are shown for judges ranked and selected based on performance over seven levels of history. The lowest curve in each graph (History 0) corresponds to choosing k judges at random, and the highest curve (History) corresponds to choosing k judges according to their true skill based on a full history

***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 (discrete)



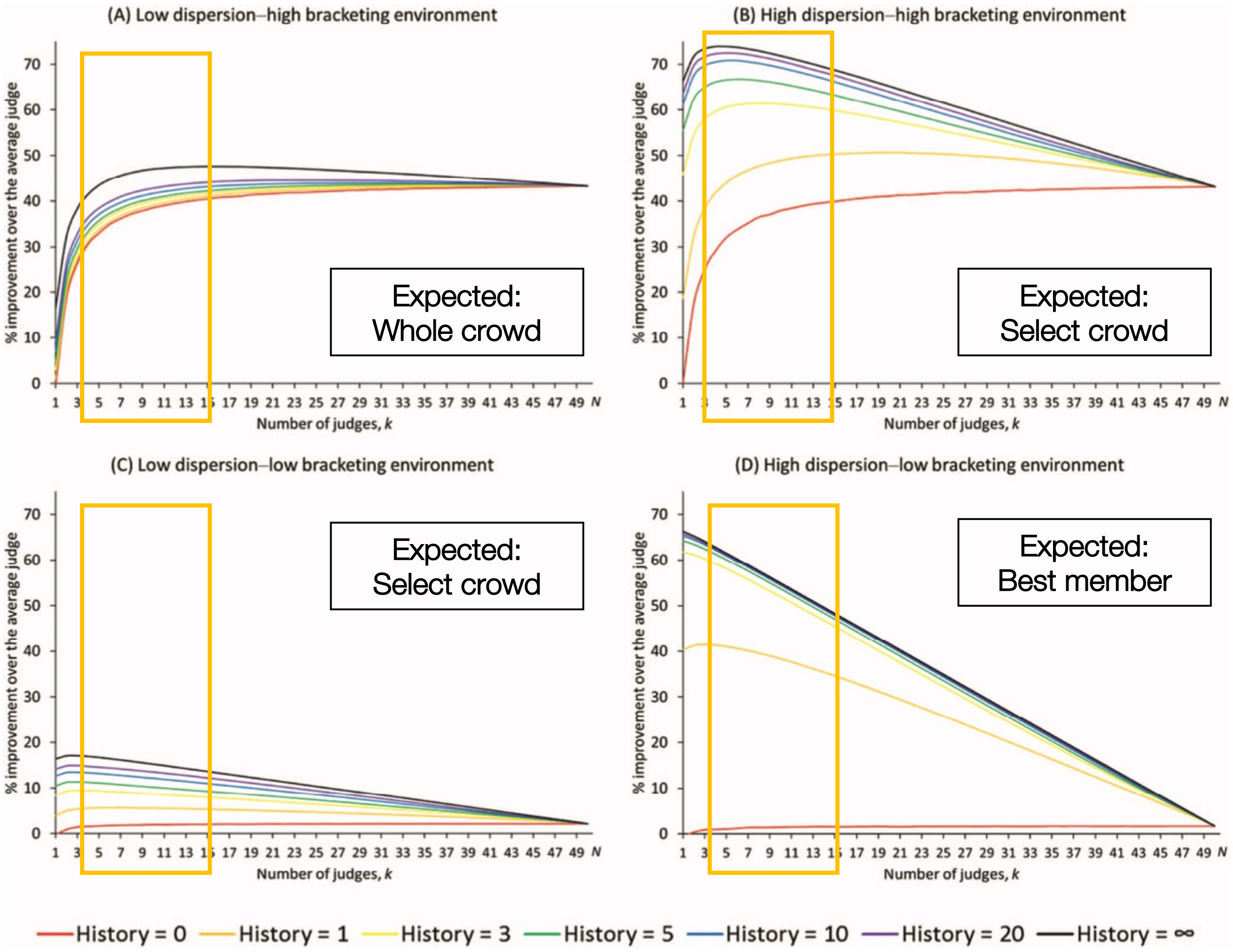
Important patterns:

1. Effect of environment on best strategy

Figure 2. Performance of judgment strategies for a simulated crowd of 50 judges. The performance of the best member is indicated at k 1, of the whole crowd at k N, and of select crowds at 1 k N. Curves are shown for judges ranked and selected based on performance over seven levels of history. The lowest curve in each graph (History 0) corresponds to choosing k judges at random, and the highest curve (History) corresponds to choosing k judges according to their true skill based on a full history

***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 (discrete)



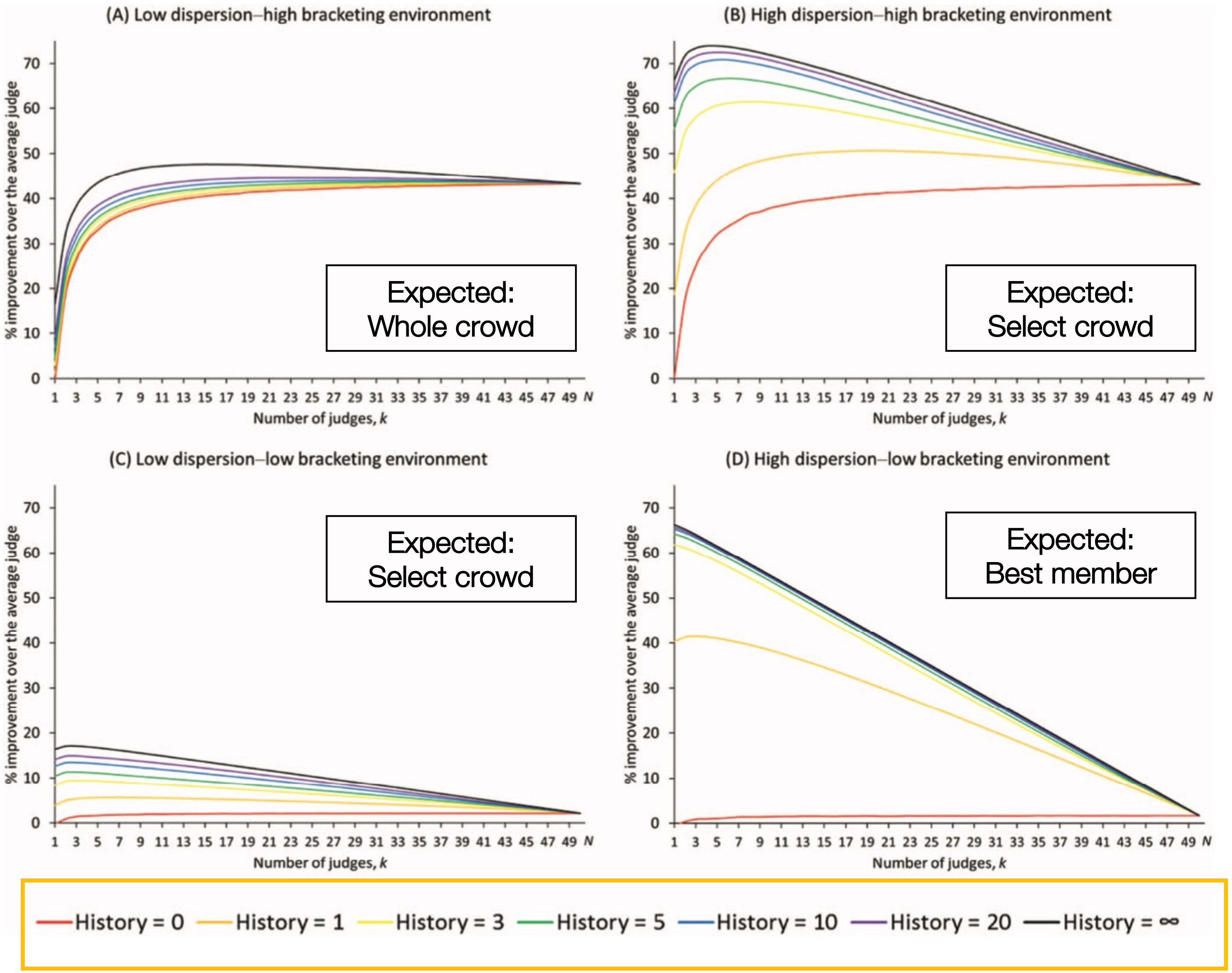
Important patterns:

1. Effect of environment on best strategy
2. Similar performance of select crowds for k +/- 5 judges

Figure 2. Performance of judgment strategies for a simulated crowd of 50 judges. The performance of the best member is indicated at k 1, of the whole crowd at k N, and of select crowds at 1 k N. Curves are shown for judges ranked and selected based on performance over seven levels of history. The lowest curve in each graph (History 0) corresponds to choosing k judges at random, and the highest curve (History) corresponds to choosing k judges according to their true skill based on a full history

***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 (discrete)



Important patterns:

1. Effect of environment on best strategy
2. Similar performance of select crowds for k +/- 5 judges
3. Performance better with longer histories (but: diminishing returns!)

Figure 2. Performance of judgment strategies for a simulated crowd of 50 judges. The performance of the best member is indicated at k 1, of the whole crowd at k N, and of select crowds at 1 k N. Curves are shown for judges ranked and selected based on performance over seven levels of history. The lowest curve in each graph (History 0) corresponds to choosing k judges at random, and the highest curve (History) corresponds to choosing k judges according to their true skill based on a full history

***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 (continuous)

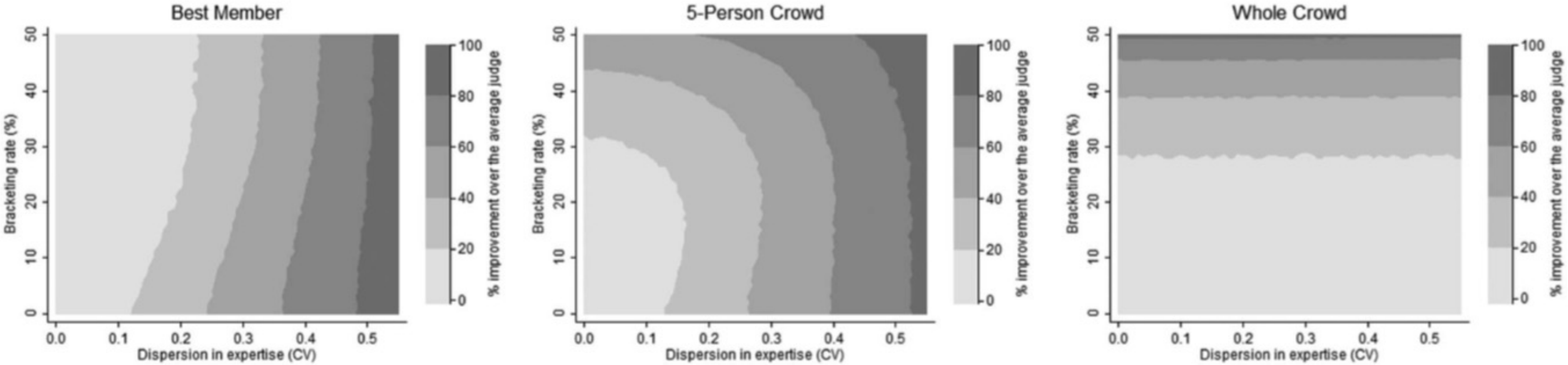


Figure 3. Contour maps of performance across 2,856 simulated judgment environments for three judgment strategies. Five trials of history were used to rank and select judges (N = 50). Darker shades of gray indicate greater percent improvement over the average judge. CV = coefficient of variation.

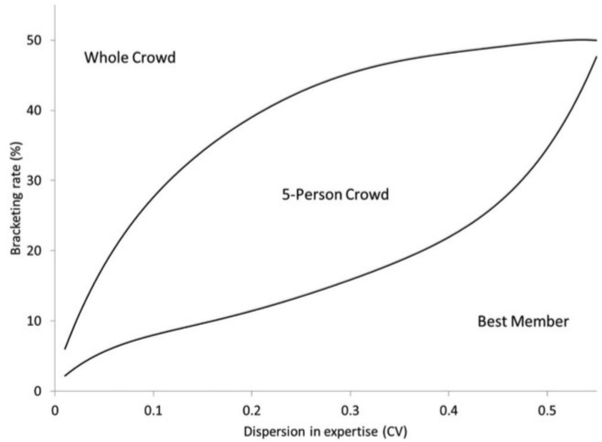


Figure 4. Best-performing strategy for each simulated judgment environment with N = 50 judges ranked and selected based on five periods of history. With less (more) history available to select judges, the curves rotate clockwise (counterclockwise). CV = coefficient of variation.

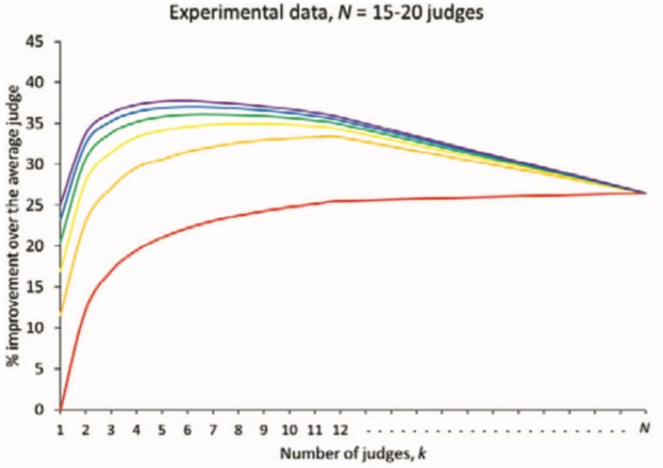
***Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology, 107*(2), 276–299.

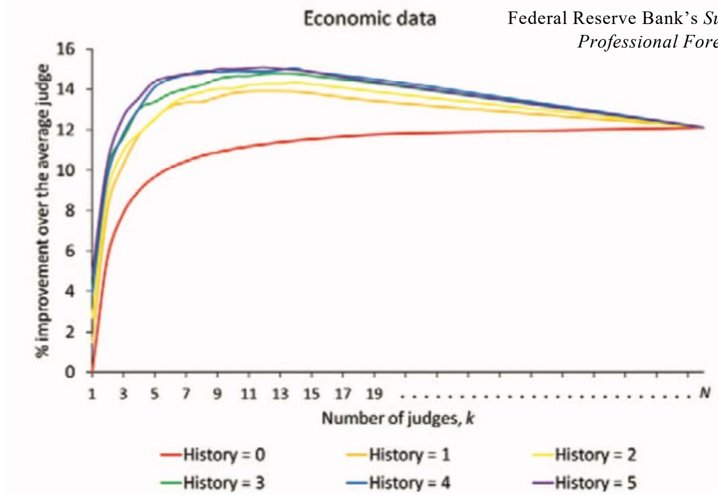
Aggregation of inferences: Real data

Table 1
Counts for Ranked Performance of the Best Member, Whole Crowd, and Select Crowd in the Experimental (N = 40) and Economic (N = 50) Data Sets

Strategy	1st	2nd	3rd
		Rank in experimental data	
Best member	5	9	26
Whole crowd	14	13	13
5-person select crowd	21	18	1
		Rank in economic data	
Best member	1	9	40
Whole crowd	15	27	8
5-person select crowd	34	14	2

Note. The best member and select crowd were ranked and selected based on five periods of history.





Federal Reserve Bank's *Survey of Professional Forecasters*

***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

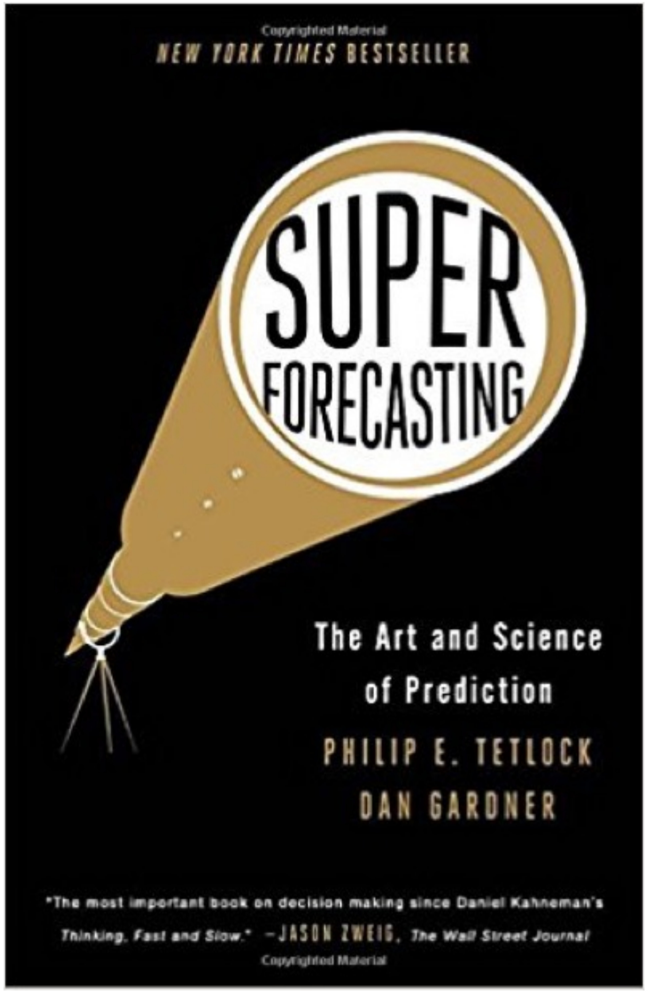
Strategy	M	SD	Difference in means: 1	2	3	4	5
1. Random economist	3.24	1.37	—				
2. Average of all economists	4.71	1.22	1.46***	—			
3. Most accurate economist last year	4.60	1.28	1.35***	−0.11	—		
4. Most accurate economist last 5 years	$\boxed{5.04}$	1.22	1.79***	0.33***	0.44***	—	
5. Average of 5 most accurate economists last year	$\boxed{5.11}$	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



https://goodjudgment.com

Welcome to

Good Judgment® Open

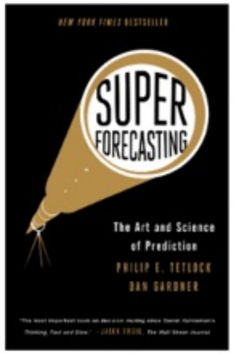
Are you a Superforecaster®?

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.

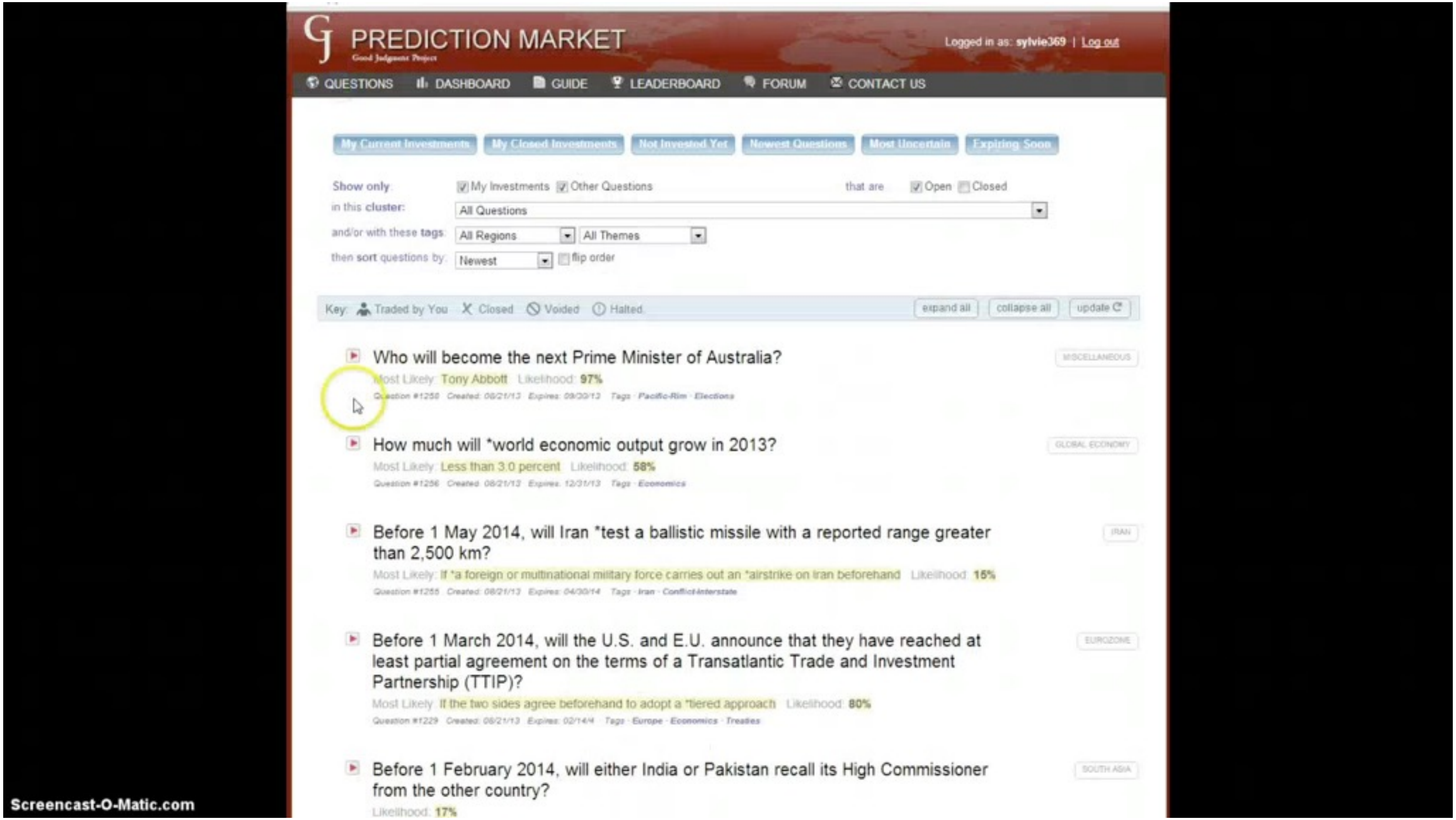
Good Judgment's co-founder, Philip Tetlock, literally wrote the book on state-of-the-art crowd-sourced forecasting. Learn more about Good Judgment and the services it provides at goodjudgment.com.

GOOD JUDGMENT

A quick peek at what the Superforecasters are saying today...

How many deaths attributed to H5N1 avian influenza will the World Health Organization (WHO) report between 7 February 2023 and 31 December 2024?	Today's Forecast	1-week Change
A Fewer than 100	100%	0
B Between 100 and 1,000, inclusive	0%	0
C More than 1,000 but fewer than 10,000	0%	0
D Between 10,000 and 100,000, inclusive	0%	0
E More than 100,000	0%	0

Good judgment project



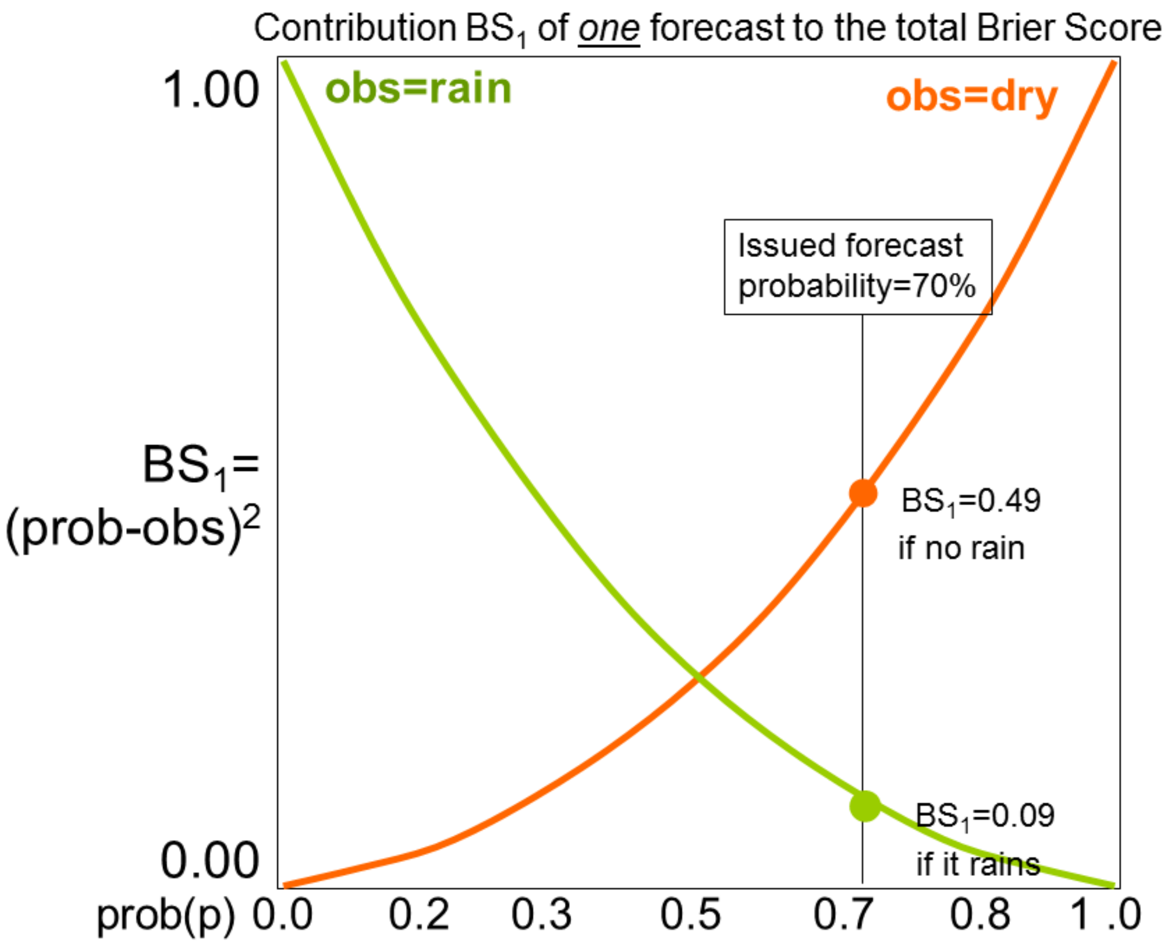
Good judgment project: Psychological interventions

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: A proper scoring rule



Brier Score (BS)

- a way to measure the accuracy of probabilistic predictions
- the lower the BS, the higher the accuracy
  - **e.g., you predict the chance of rain tomorrow to be 70%.**
  - **if it rains, your Brier score is (0.7 - 1)^2 = 0.09.**
  - **if it does not rain, your Brier score is (0.7 - 0)^2 = 0.49.**
- ranges between 0 and 1

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

Good judgment project: Psychological interventions

Winning a Geopolitical Forecasting Tournament 1109

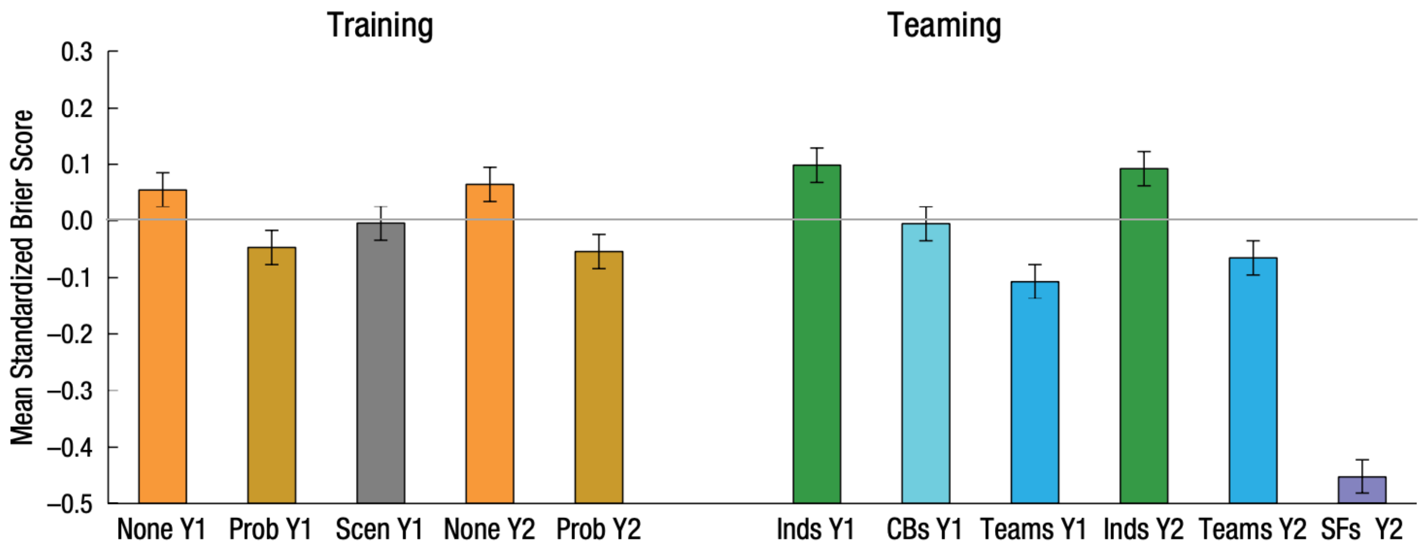


Fig. 1. Effects of training, teaming, and tracking on average Brier scores in Year 1 (Y1) and Year (Y2). The bars at the left show results for the no-training ("None"), probability-training ("Prob"), and scenario-training ("Scen") conditions; the bars at the right show results for independent forecasters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

Check your understanding:

If BS ranges between 0 and 1, and lower BS means higher accuracy, what does a negative mean standardized BS tell you about the impact of training versus teaming and tracking?

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

Mellers, B. A. & Tetlock, P. E. (2019). From discipline-centered rivalries to solution-centered science. *American Psychologist*, *74*(3), 290-300. http://doi: 10.1037/amp0000429

Your turn!

Image created with AI (Bing), February 13, 2024

What do you think makes a superforecaster?

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.

"[…] superforecasters have distinctive dispositional profiles, scoring higher on several measures of fluid intelligence and crystallized intelligence, higher on the desire to be the best, the need for cognition, open-minded thinking, and endorsements of a scientific worldview with little tolerance for supernaturalism. Table 3 shows that these same variables correlate with forecasting accuracy."

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

Table 3. Correlates With Measures With Accuracy

Measure	Correlation	$t(1774)$	p
[illegible] Progressive Matrices	–.18	–7.70	<.001
[illegible]	–.22	–9.49	<.001
[illegible]	–.09	–3.80	<.001
[illegible]	–.16	–6.82	<.001
[illegible]	–.23	–9.95	<.001
[illegible]	–.16	–6.82	<.001
[illegible]	–.12	–5.09	<.001
[illegible]	–.18	–7.70	<.001
[illegible]	–.14	–5.95	<.001
[illegible]	–.11	–4.66	<.001
[illegible]	–.07	–2.95	<.002
[illegible]	–.12	–5.09	<.001
[illegible] checked	–.18	–7.70	<.001
Average [illegible] 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.

But wait … what do you make of the size of these correlations?

"[…] superforecasters have distinctive dispositional profiles, scoring higher on several measures of fluid intelligence and crystallized intelligence, higher on the desire to be the best, the need for cognition, open-minded thinking, and endorsements of a scientific worldview with little tolerance for supernaturalism. Table 3 shows that these same variables correlate with forecasting accuracy."

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

Summary

- **Statisticized groups:** Statisticized 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.

- **Crowds vs. single experts:** 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!

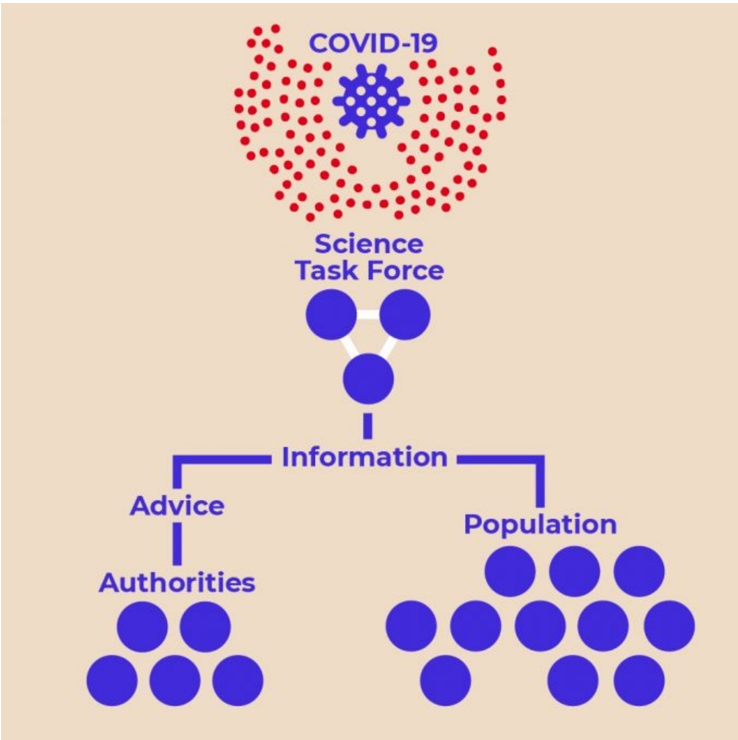
- **Select crowds:** Often, teams of experts seem to provide a good balance by capitalising on dispersion and bracketing.

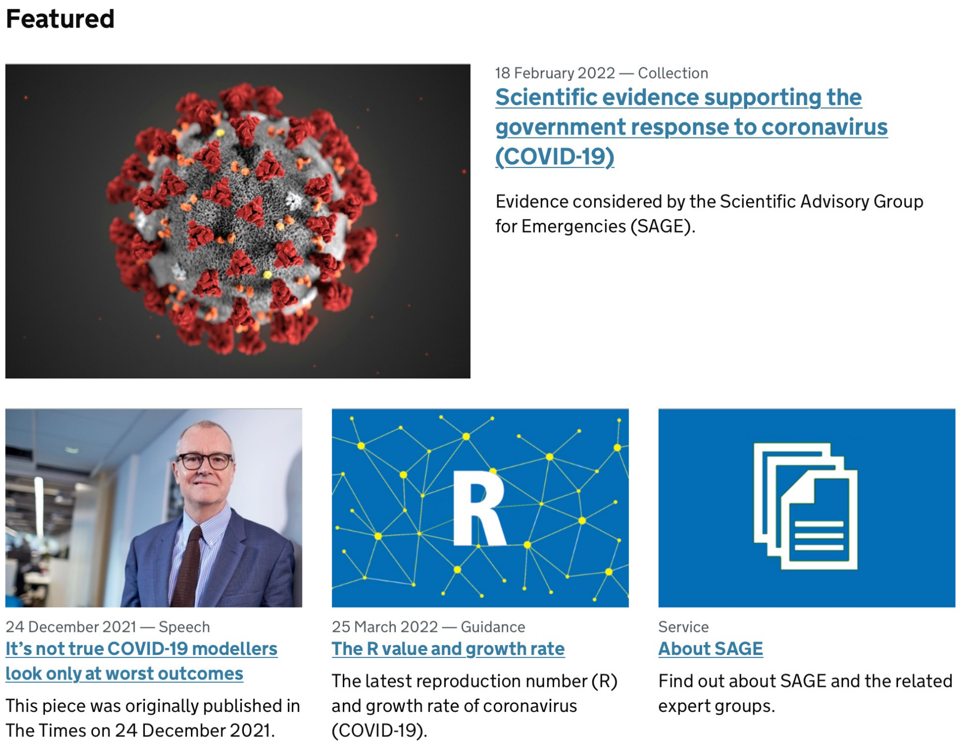
- **Psychological interventions:** Training, teaming, and tracking (processes which incorporate probability training, scenario thinking, and forecast averaging) can meaningfully enhance judgment and improve forecasting accuracy.

Have a good week and see you in April!

Appendix (not mandatory)
Example: Improving science task forces

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/

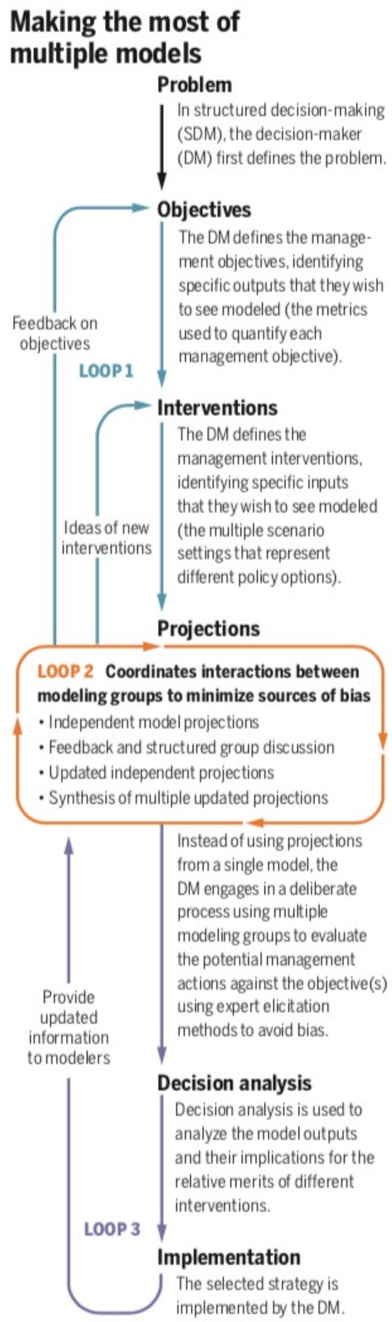
https://www.youtube.com/watch?v=L7uBwyr0sdg

Appendix (not mandatory)
Example: Improving science task forces

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

Appendix (not mandatory)
Example: Improving peer review

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.

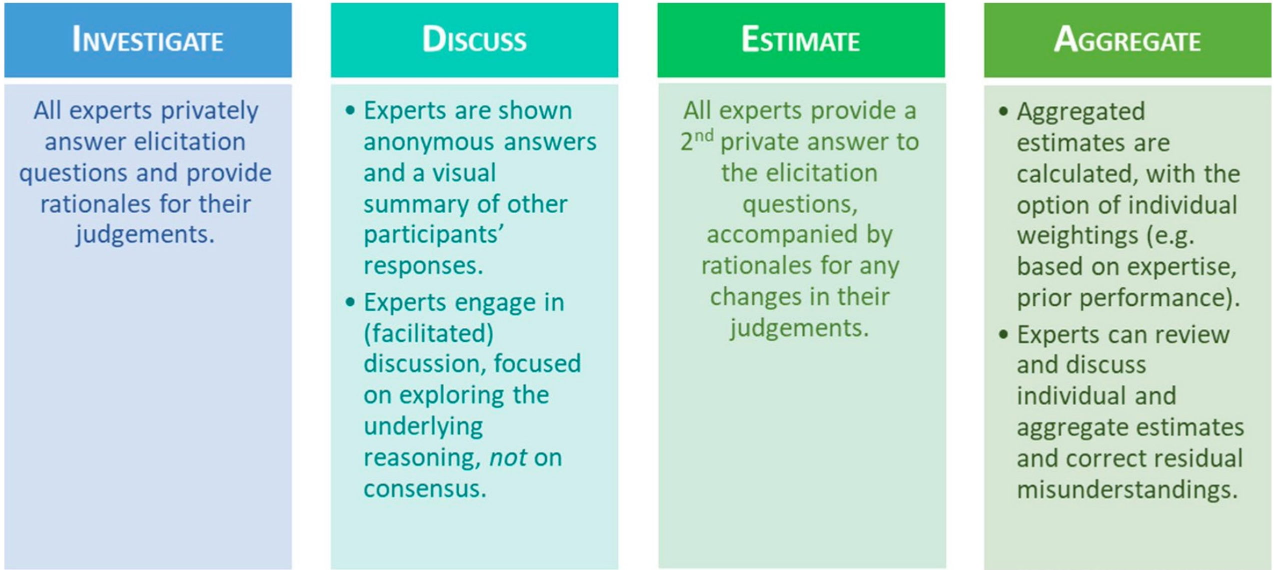


Fig. 1 The IDEA protocol for structured expert judgement elicitation (adapted from [20])

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

Appendix (not mandatory)
A better crystal ball: The inner crowd

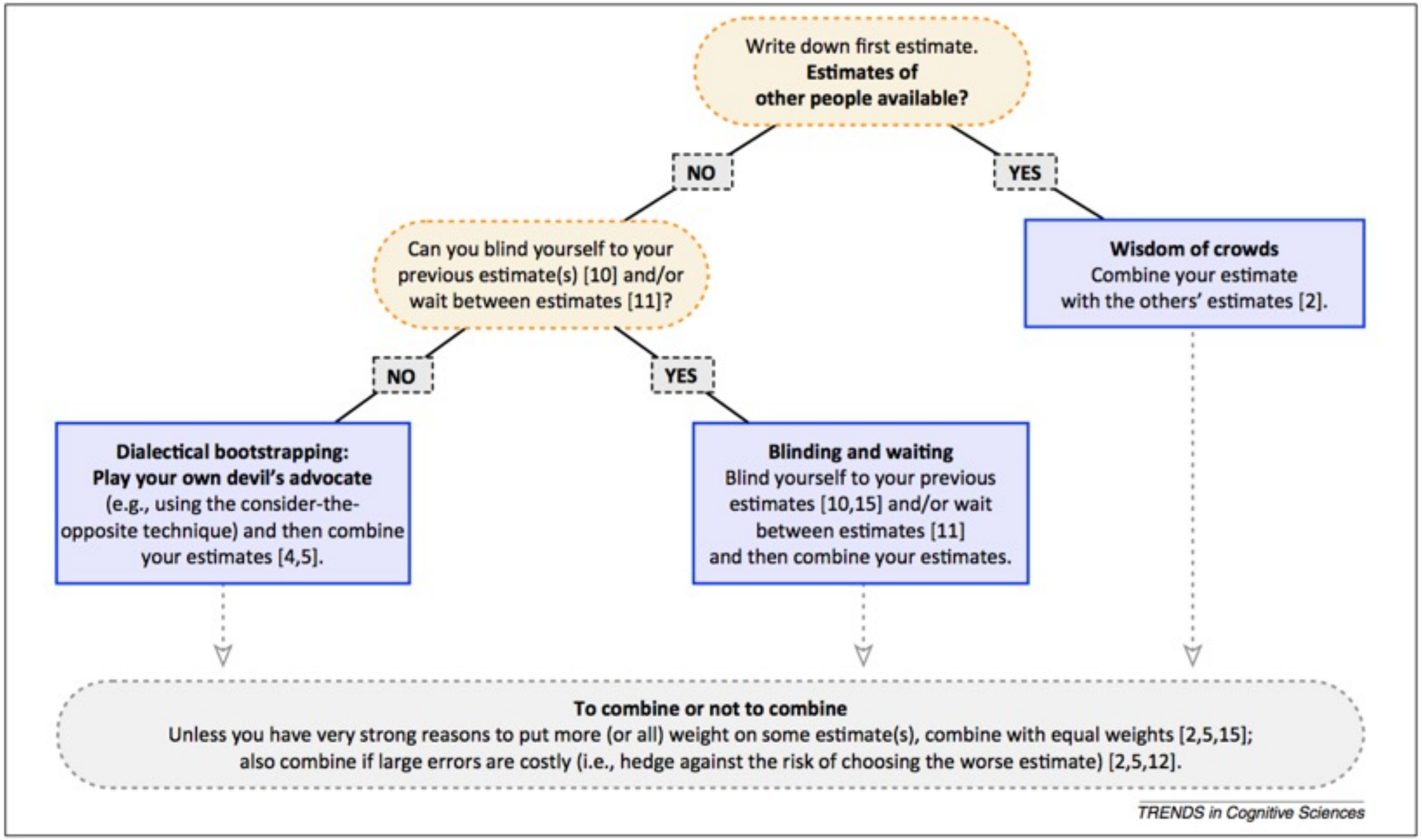


Figure 1. Decision tree for deciding when and how to use the inner crowd.

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