

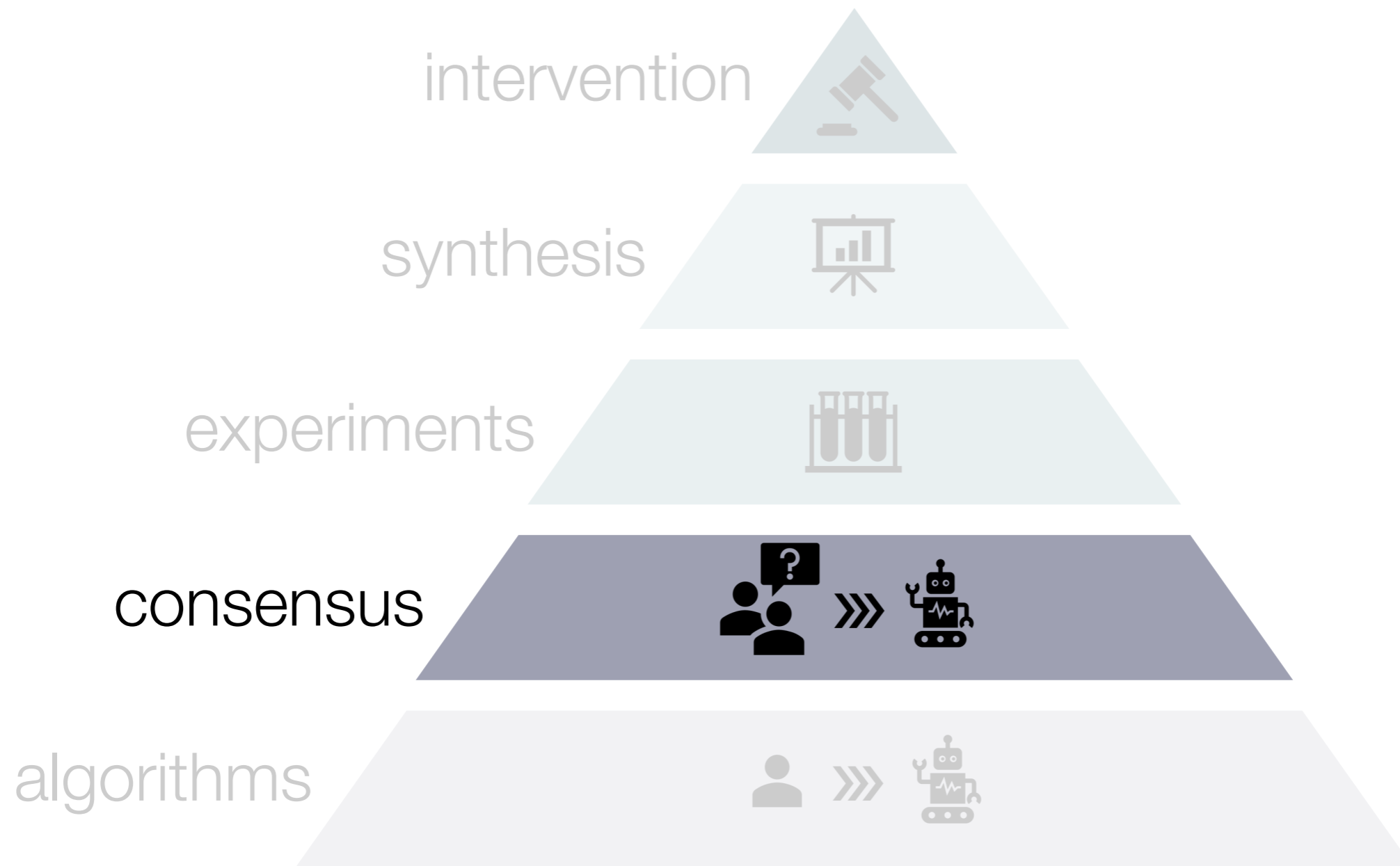
Evidence-based Decision Making Consensus: The wisdom of experts

Loreen Tisdall, FS 2024

Version: March 24, 2024

Question on relevance

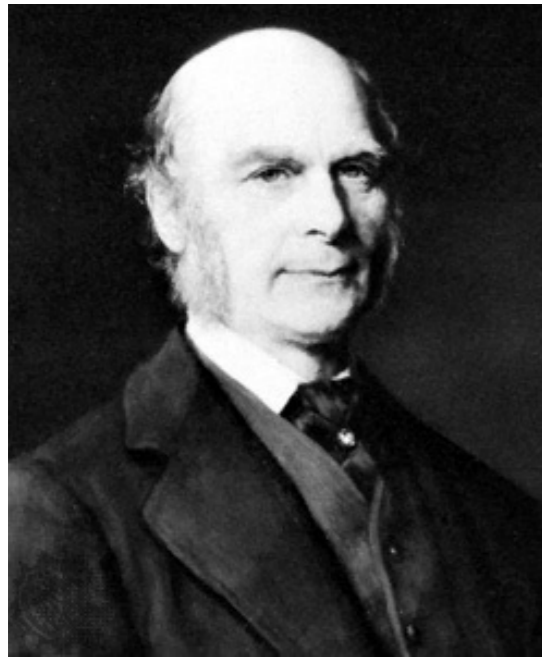
What **DO YOU** think is the relevance of evidence-based decision making for your studies, your career path, your personal development, etc.?



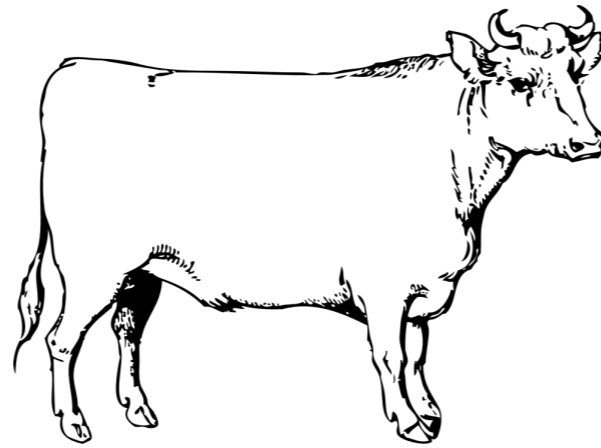
Goals for today

- Understand the performance of groups as a process of statistical aggregation and learn about how to predict 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.
- Debate possible implications for application to societal issues

When groups work: Wisdom of the crowd!



Francis Galton, 1822-1911



True = 1198 pounds

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		Excess of Observed over Normal
		Observed deviates from 1207 lbs.	Normal p.e = 37	
5	1074	- 133	- 90	+ 43
10	1109	- 98	- 70	+ 28
15	1126	- 81	- 57	+ 24
20	1148	- 59	- 46	+ 13
<i>q</i> 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
<i>m</i> 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
<i>q</i> 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.

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

Staticized groups can be powerful!

**A BIOLOGIST, A CHEMIST, AND
A STATISTICIAN ARE OUT HUNTING.**

The biologist shoots at a deer and misses 5 feet to the left.

The chemist takes a shot and misses 5 feet to the right.

The statistician yells "We got 'em!"

Not just your average kind of joke :)

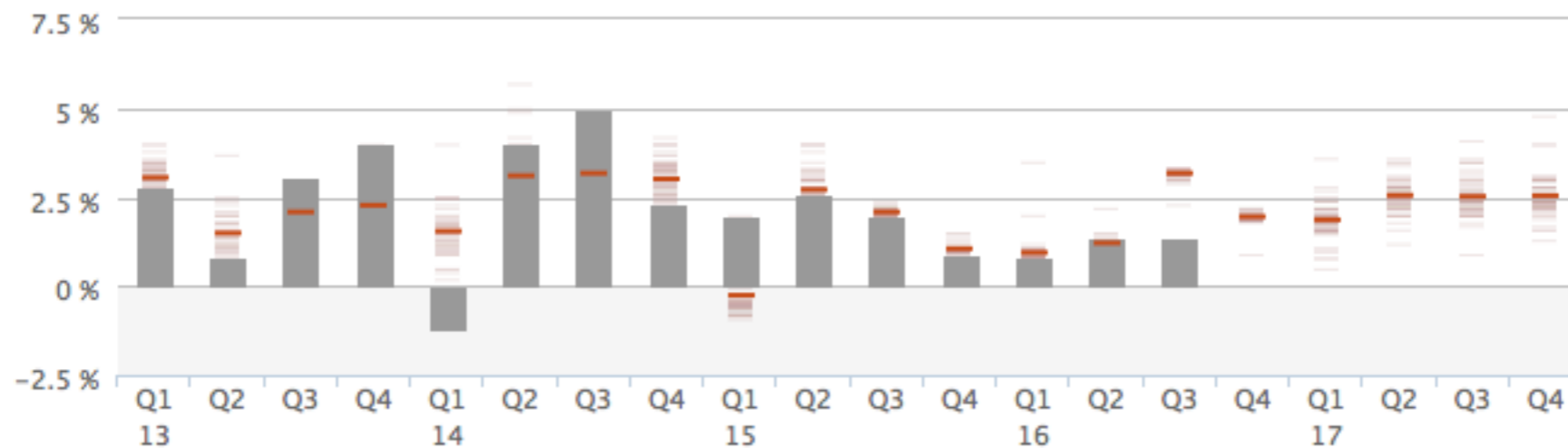
THE WALL STREET JOURNAL.

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)

1.4%

Projected: Q4 2016

1.9% ▲

Projected: Q1 2017

1.9%

Projected: Q2 2017

2.5%

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)

1.4%



Projected: Q4 2016

1.9% ▲



Projected: Q1 2017

1.9%



Projected: Q2 2017

2.5%



The wisdom of *select* crowds

In this paper, Mannes and colleagues:

1. use simulations to show the relative performance of crowds, best judge, or select crowds as a function of environment/judge performance
2. show the relative performance of crowds, best judge, or select crowds in real environments
3. use surveys/experiments to evaluate people's intuitions about the performance of staticized groups (crowds, select crowds) vs. best judge

Aggregation of inferences

Expect that 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)

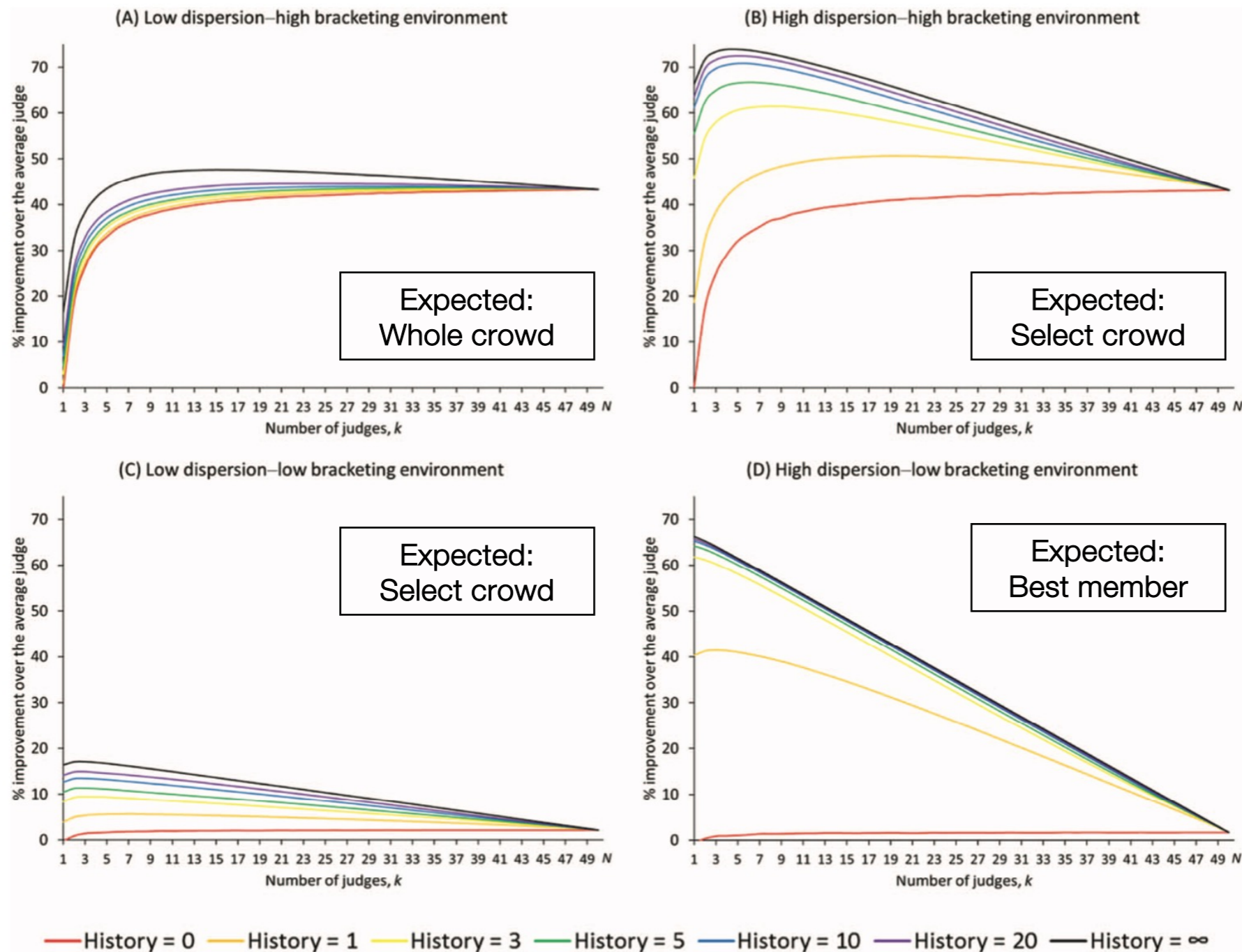
- **Dispersion in expertise:** degree to which members differ in ability to estimate the criterion, regardless of the level of expertise (e.g., zero dispersion could be all novices or all experts)
- **Bracketing:** frequency with which any two judges fall on opposite (either) sides of the criterion

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

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

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

Aggregation of inferences: Simulations (discrete)



Important patterns:

1. Effect of environment on best strategy
2. Similar performance of select crowds for $k \pm 5$ 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

Aggregation of inferences: Simulations (continuous)

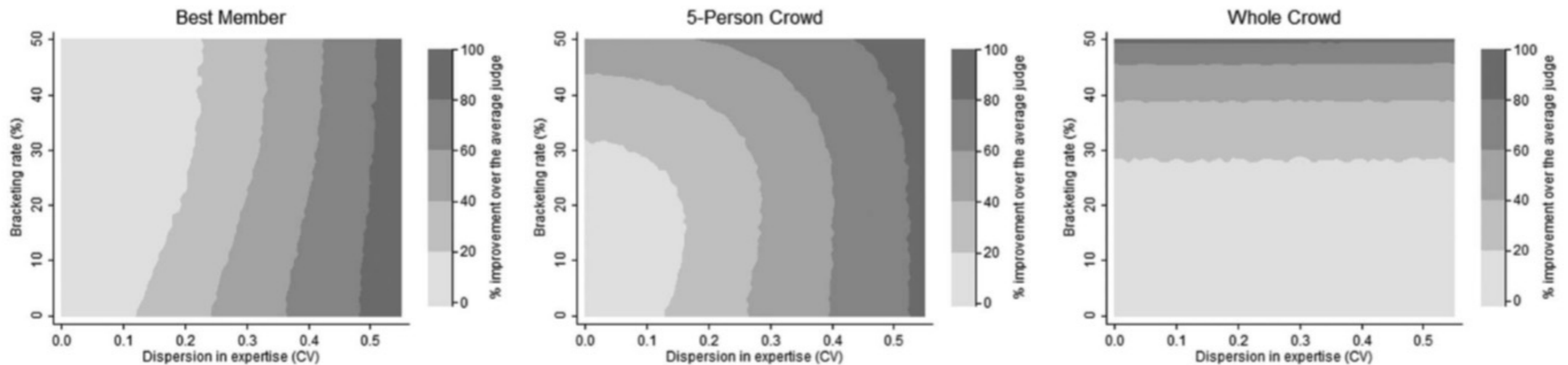


Figure 3. Contour maps of performance across 2,856 simulated judgment environments for three judgment strategies. Five trials of history were used to rank and select judges ($N = 50$). Darker shades of gray indicate greater percent improvement over the average judge. CV = coefficient of variation.

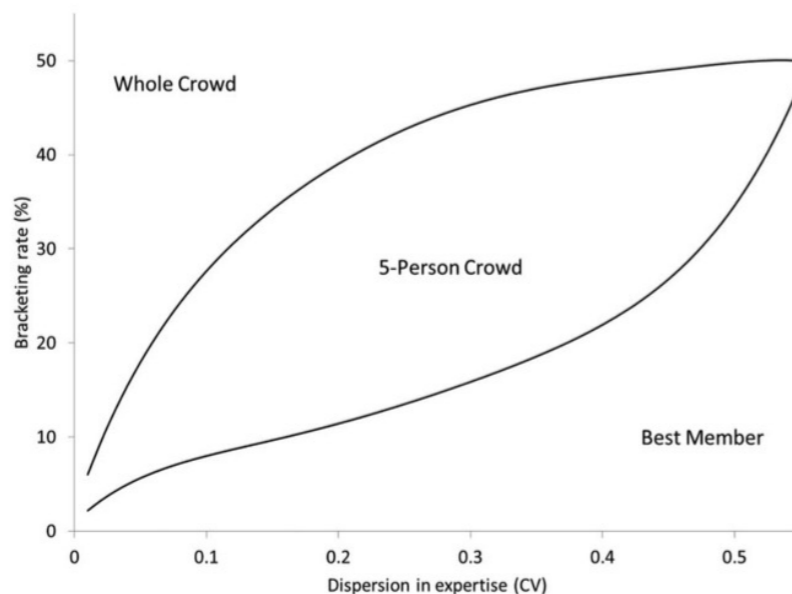


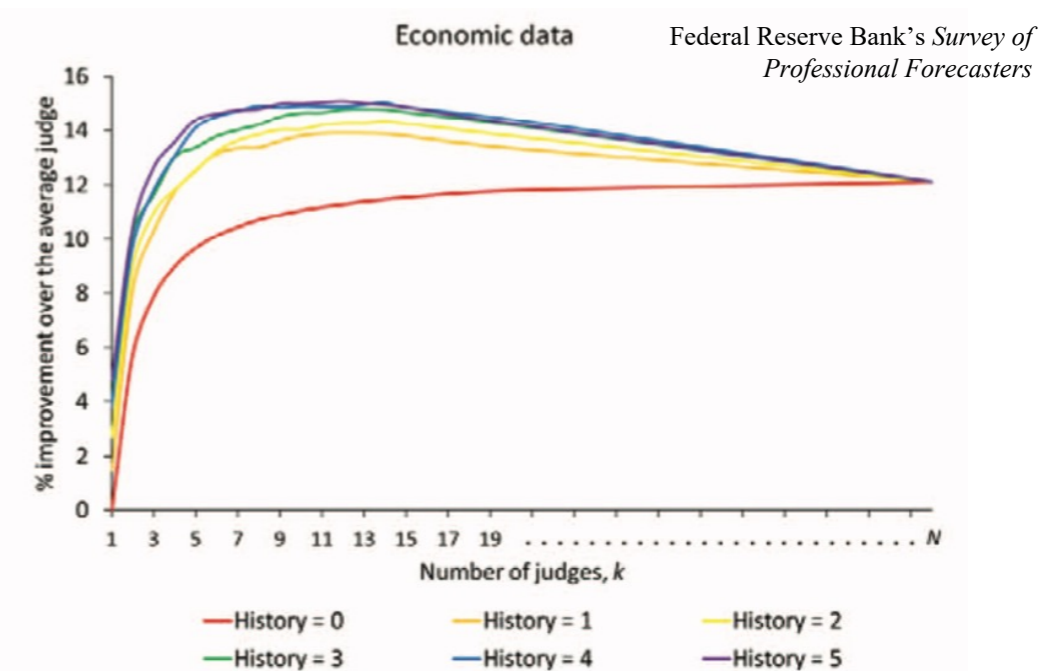
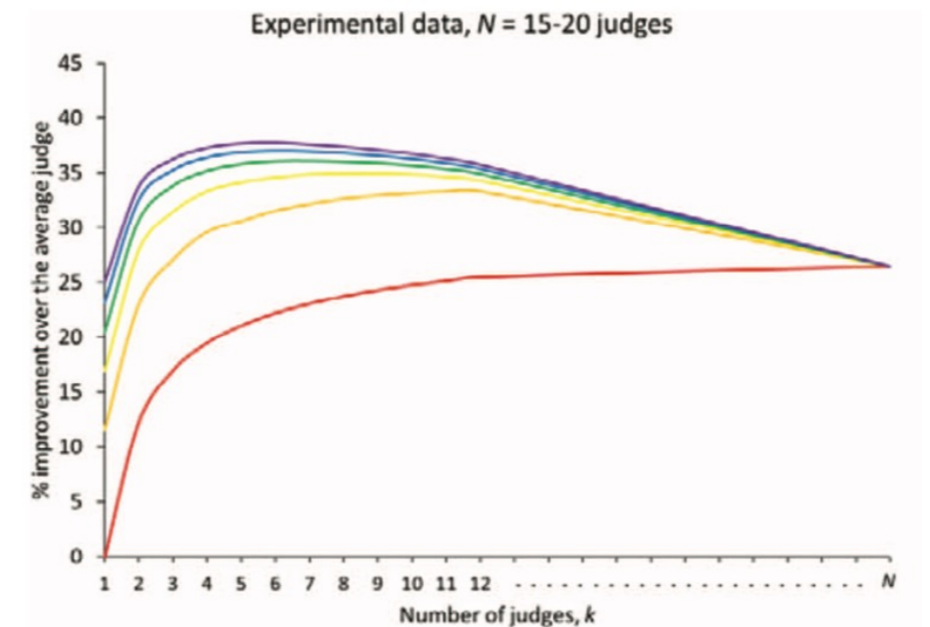
Figure 4. Best-performing strategy for each simulated judgment environment with $N = 50$ judges ranked and selected based on five periods of history. With less (more) history available to select judges, the curves rotate clockwise (counterclockwise). CV = coefficient of variation.

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.



Aggregation of inferences: Lay intuitions

Table 2

Ratings of Judgment Strategies in Experiment 1

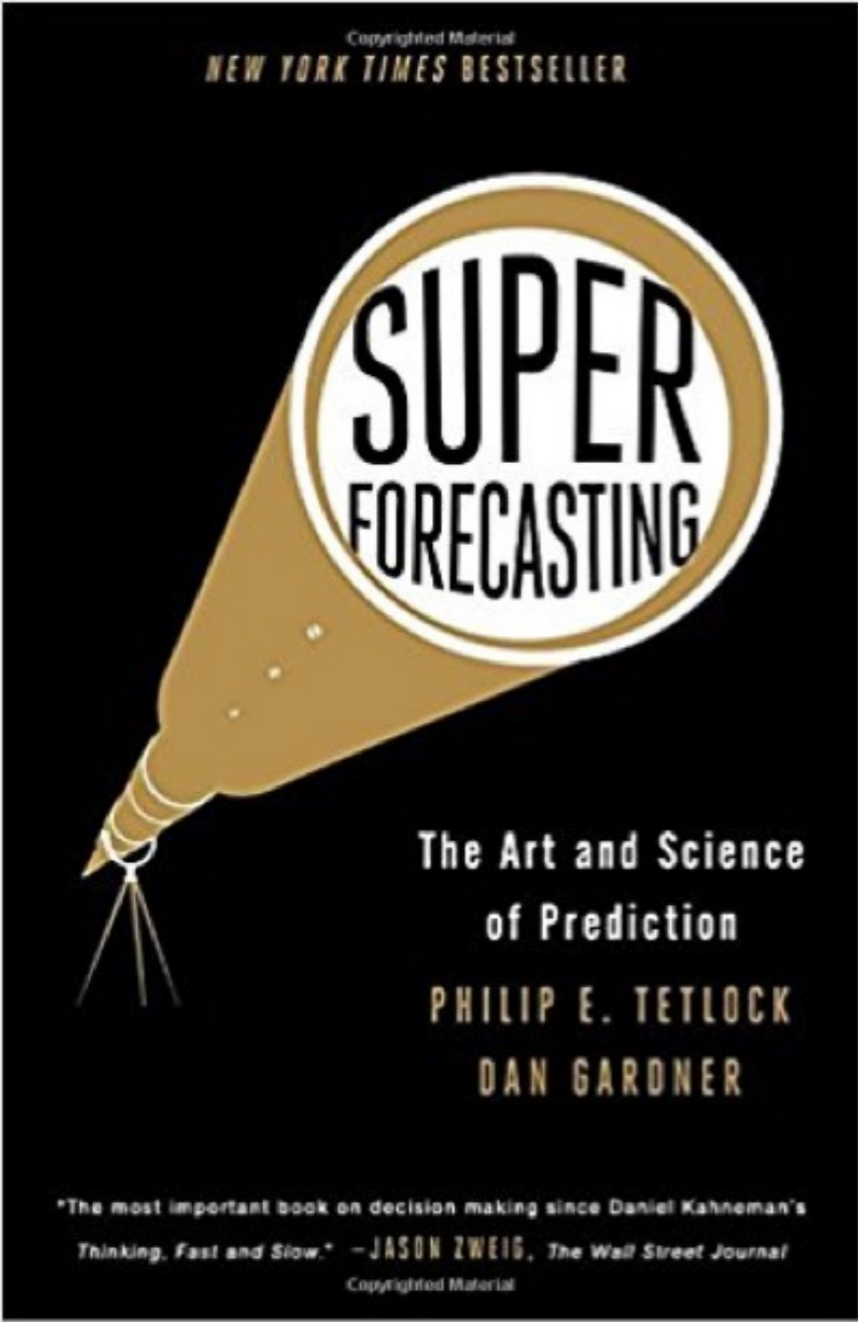
Strategy	<i>M</i>	<i>SD</i>	Difference in means					
			1	2	3	4	5	
1. Random economist	3.24	1.37	—					
2. Average of all economists	4.71	1.22	1.46***	—				
3. Most accurate economist last year	4.60	1.28	1.35***	-0.11	—			
4. Most accurate economist last 5 years	5.04	1.22	1.79***	0.33***	0.44***	—		
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.

Good Judgment Project



Welcome to Good Judgment® Open

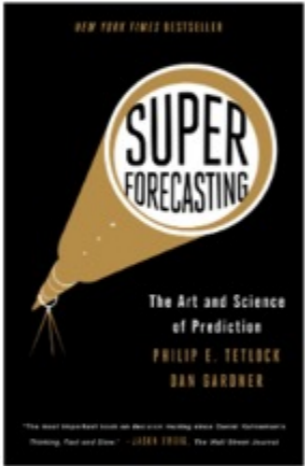
Are you a Superforecaster®?

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.



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](#).



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

<https://goodjudgment.com>

Good Judgment Project

PREDICTION MARKET
Good Judgment Project

Logged in as: [sylvie369](#) | [Log out](#)

QUESTIONS | DASHBOARD | GUIDE | LEADERBOARD | FORUM | CONTACT US

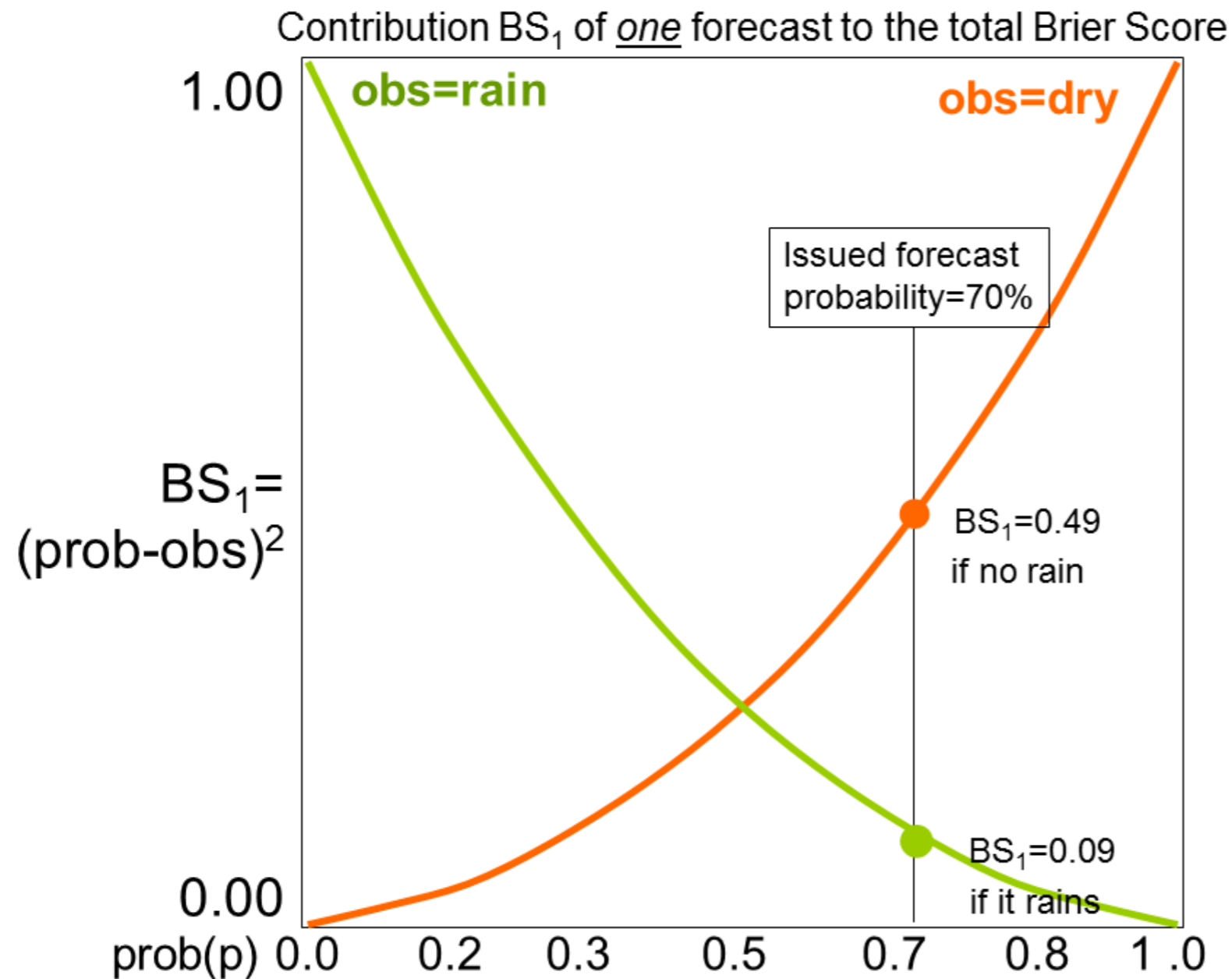
[My Current Investments](#) | [My Closed Investments](#) | [Not Invested Yet](#) | [Newest Questions](#) | [Most Uncertain](#) | [Expiring Soon](#)

Show only: My Investments Other Questions that are Open Closed
in this cluster:
and/or with these tags:
then sort questions by: flip order

Key: Traded by You Closed Voided Halted [expand all](#) [collapse all](#) [update](#)

- Who will become the next Prime Minister of Australia?** MISCELLANEOUS
Most Likely: **Tony Abbott** Likelihood: **97%**
Question #1250 Created: 08/21/13 Expires: 09/30/13 Tags: Pacific-Rim - Elections
- How much will *world economic output grow in 2013?** GLOBAL ECONOMY
Most Likely: **Less than 3.0 percent** Likelihood: **58%**
Question #1256 Created: 08/21/13 Expires: 12/31/13 Tags: Economics
- Before 1 May 2014, will Iran *test a ballistic missile with a reported range greater than 2,500 km?** IRAN
Most Likely: **If *a foreign or multinational military force carries out an *airstrike on Iran beforehand** Likelihood: **15%**
Question #1255 Created: 08/21/13 Expires: 04/30/14 Tags: Iran - Conflict/Interstate
- Before 1 March 2014, will the U.S. and E.U. announce that they have reached at least partial agreement on the terms of a Transatlantic Trade and Investment Partnership (TTIP)?** EUROZONE
Most Likely: **If the two sides agree beforehand to adopt a *tiered approach** Likelihood: **80%**
Question #1229 Created: 08/21/13 Expires: 02/14/14 Tags: Europe - Economics - Treaties
- Before 1 February 2014, will either India or Pakistan recall its High Commissioner from the other country?** SOUTH ASIA
Likelihood: **17%**

Good Judgment Project



Brier Score (BS)

- a way to measure the accuracy of probabilistic predictions
- the lower the BS, the higher the accuracy
- ranges between 0 and 1

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: Psychological interventions

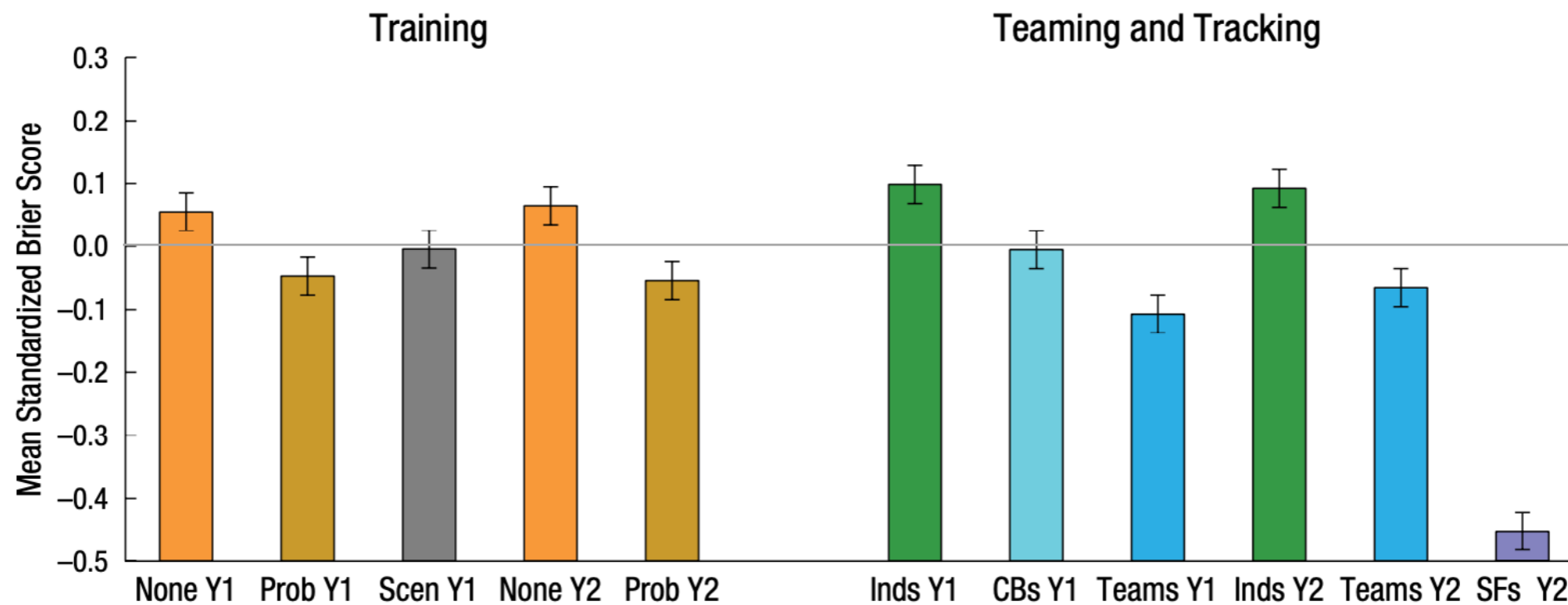


Fig. 1. Effects of training, teaming, and tracking on average Brier scores in Year 1 (Y1) and Year (Y2). The bars at the left show results for the no-training (“None”), probability-training (“Prob”), and scenario-training (“Scen”) conditions; the bars at the right show results for independent forecasters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

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>

Good Judgment Project: Superforecasters



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

Your turn!

**What do you think makes
a Superforecaster?**

Good Judgment Project: Superforecasters

Table 3. Correlates With Measures With Accuracy

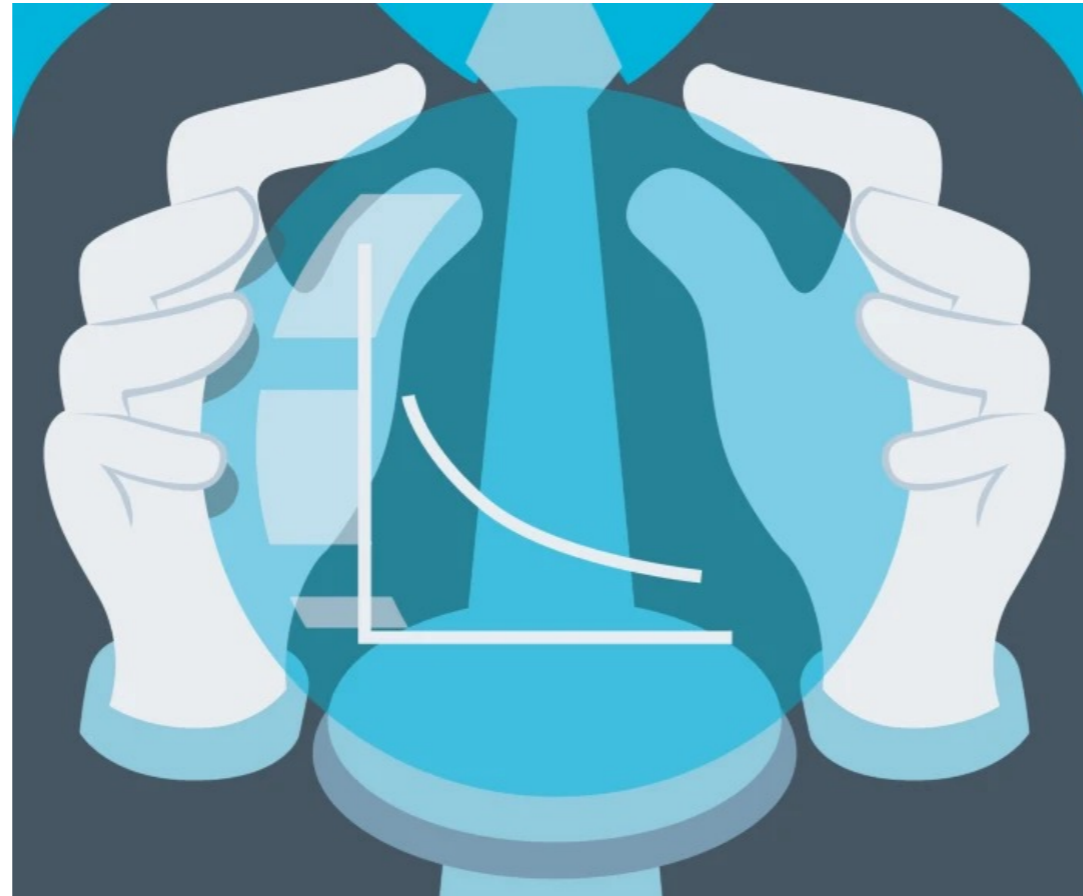
Measure	Correlation	<i>t</i> (1774)	<i>p</i>
Raven's Advanced Progressive Matrices	-.18	-7.70	<.001
Shipley-2 Abstraction Test	-.22	-9.49	<.001
Shipley-2 Vocabulary	-.09	-3.80	<.001
CRT	-.16	-6.82	<.001
Extended CRT	-.23	-9.95	<.001
Numeracy	-.16	-6.82	<.001
Political knowledge (Year 1)	-.12	-5.09	<.001
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>

A better crystal ball: Integration of approaches



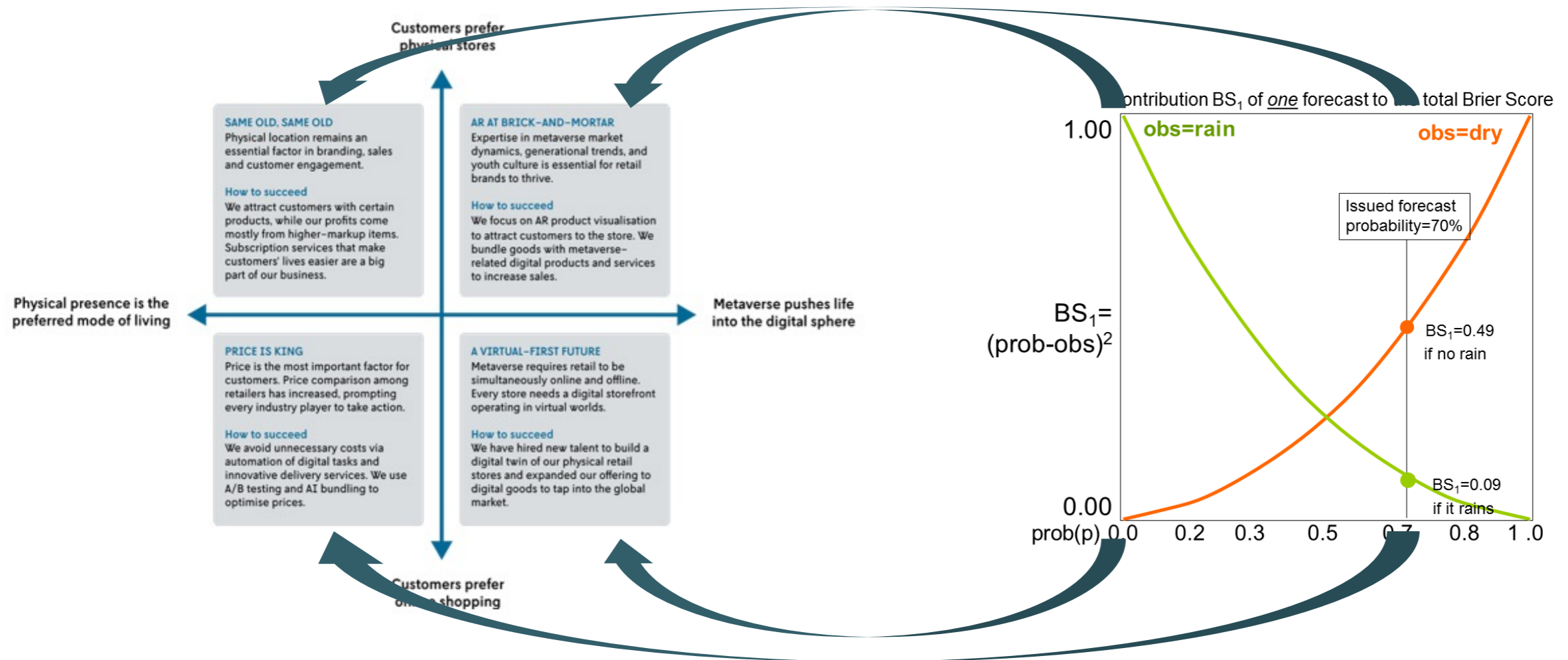
Scenario planning

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

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

A better crystal ball: The inner crowd

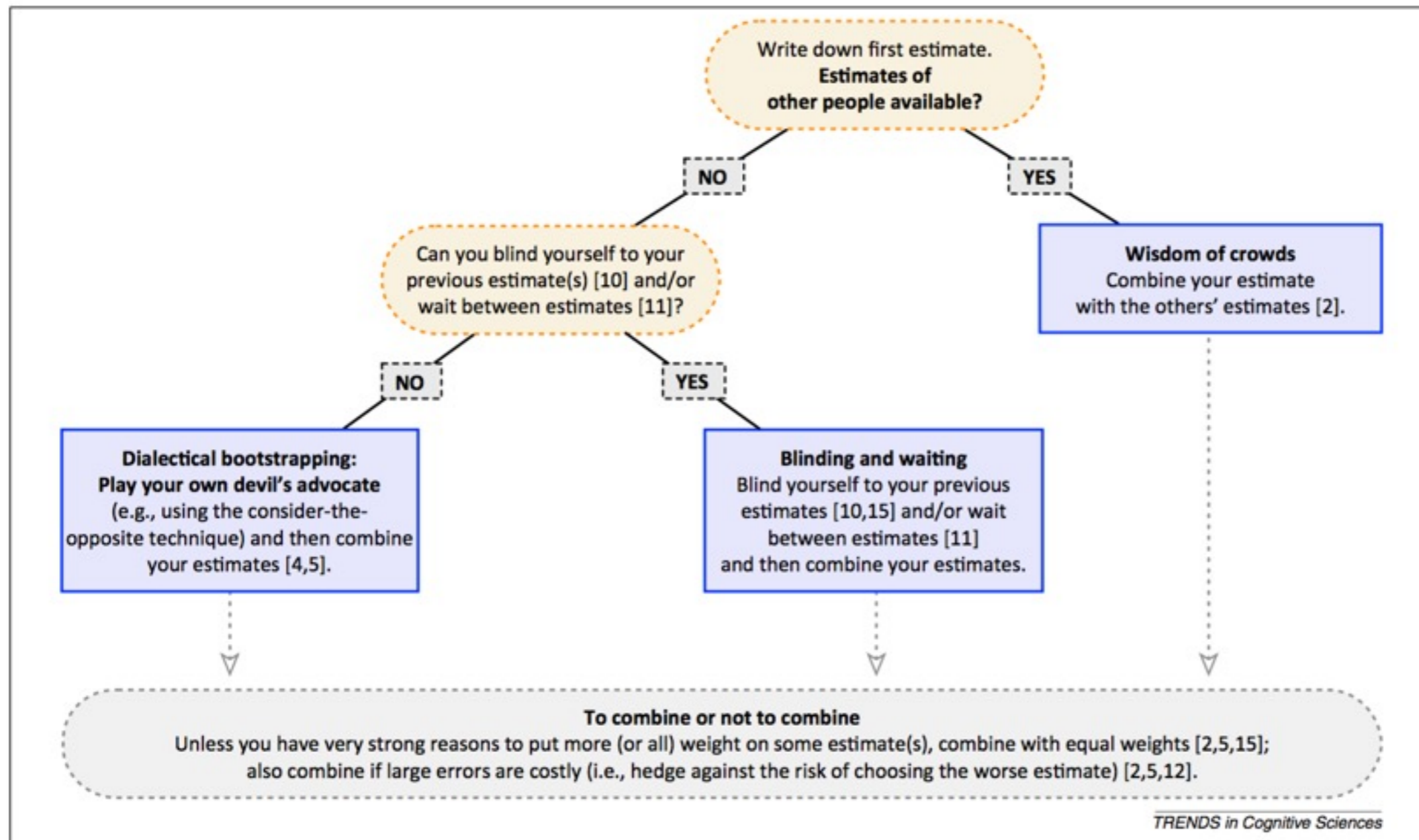


Figure 1. Decision tree for deciding when and how to use the inner crowd.

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

Implications



https://cdsbaseel.github.io/Diversity_hackathon/

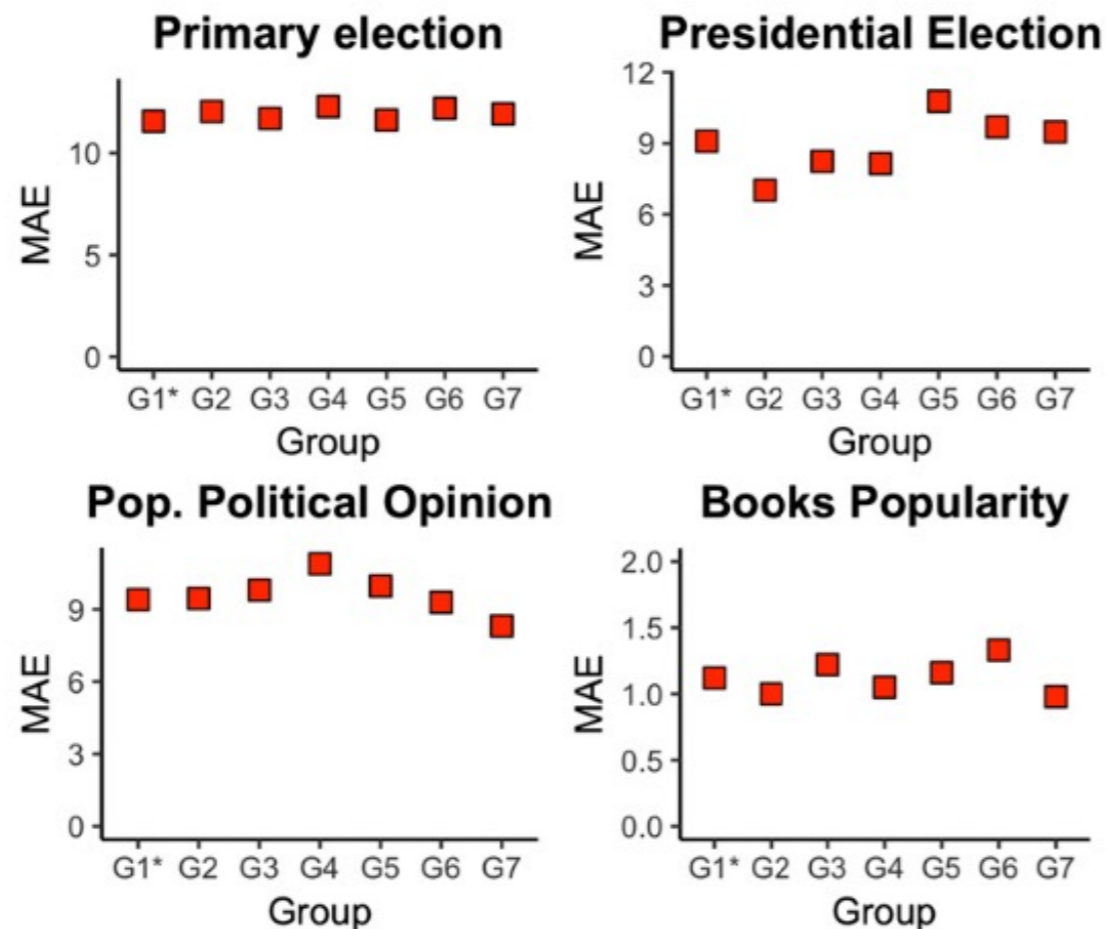
Demographically diverse crowds are not much wiser than homogeneous crowds

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.

Demographically diverse crowds are not much wiser than homogeneous crowds



Mean Absolute Error (MAE)

- measure of errors between paired observations expressing the same phenomenon
- calculated as the sum of absolute errors divided by the sample size

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.

Implications...

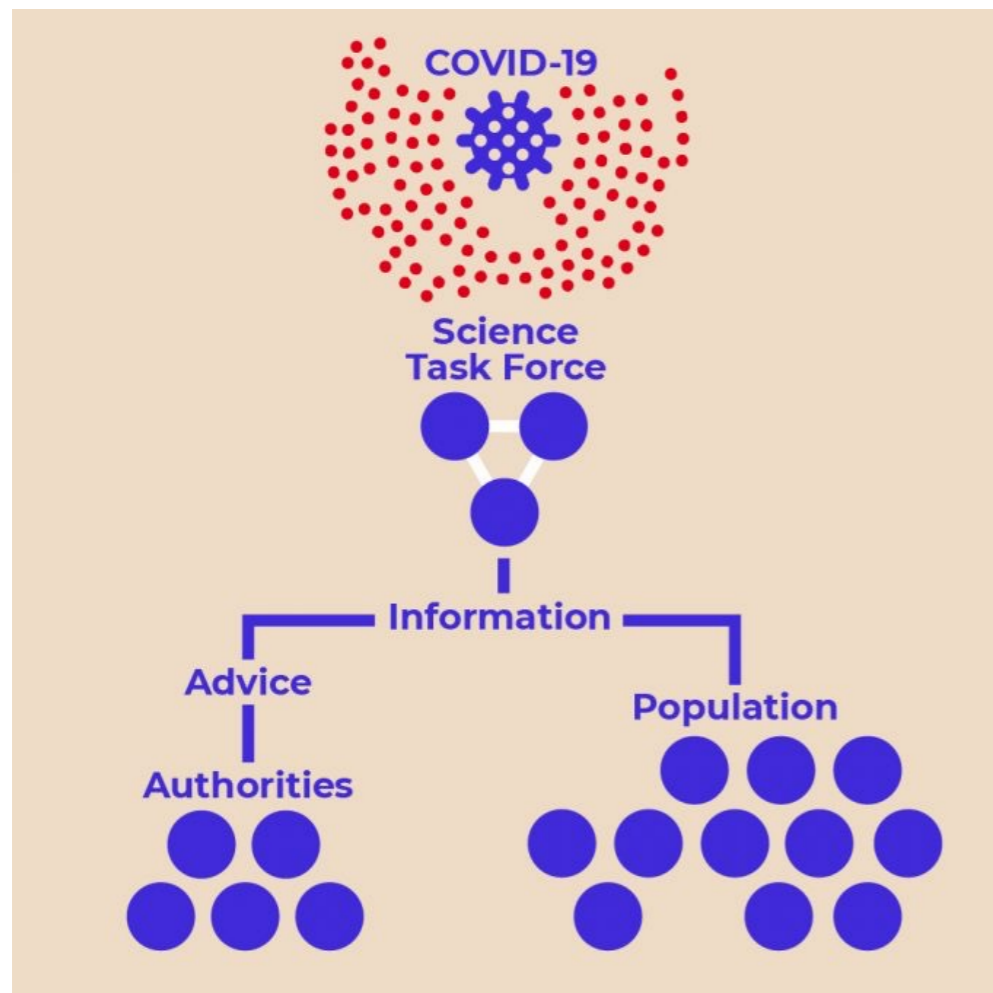
“Despite advocates’ insistence that women on boards enhance corporate performance and that diversity of task groups enhances their performance, research findings are mixed, and repeated meta-analyses have yielded average correlational findings that are null or extremely small. Therefore, social scientists 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.”

Summary

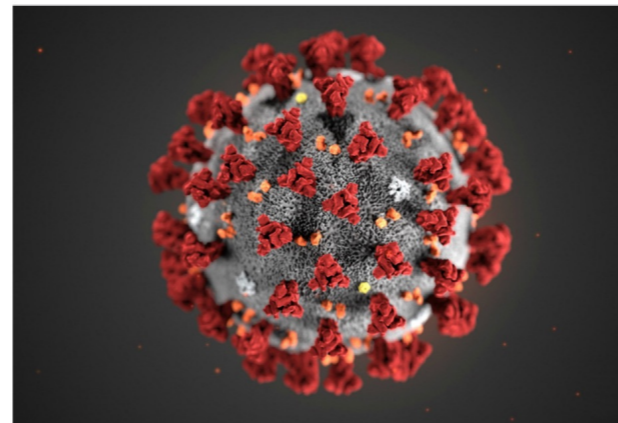
- **Staticized groups:** Staticized groups can work well. Understanding the performance of groups as a process of statistical aggregation involving different factors - dispersion and bracketing - helps predict when select crowds (or other types of aggregation) will do best.
- **Crowds vs. single experts:** 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. Lay people are not fully aware of the power of aggregation and of select crowds.
- **Implications:** Beware of drawing implications for diversity management: the literature is not yet mature but many mixed findings concerning diversity for performance. **Instead, we should argue for diversity based on ethical, not performance grounds!**

Exercise: Improving Science Task Forces

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



Featured



18 February 2022 — Collection
[Scientific evidence supporting the government response to coronavirus \(COVID-19\)](#)

Evidence considered by the Scientific Advisory Group for Emergencies (SAGE).



24 December 2021 — Speech
[It's not true COVID-19 modellers look only at worst outcomes](#)

This piece was originally published in The Times on 24 December 2021.



25 March 2022 — Guidance
[The R value and growth rate](#)

The latest reproduction number (R) and growth rate of coronavirus (COVID-19).



Service
[About SAGE](#)

Find out about SAGE and the related expert groups.

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

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

Exercise: Improving Science Task Forces

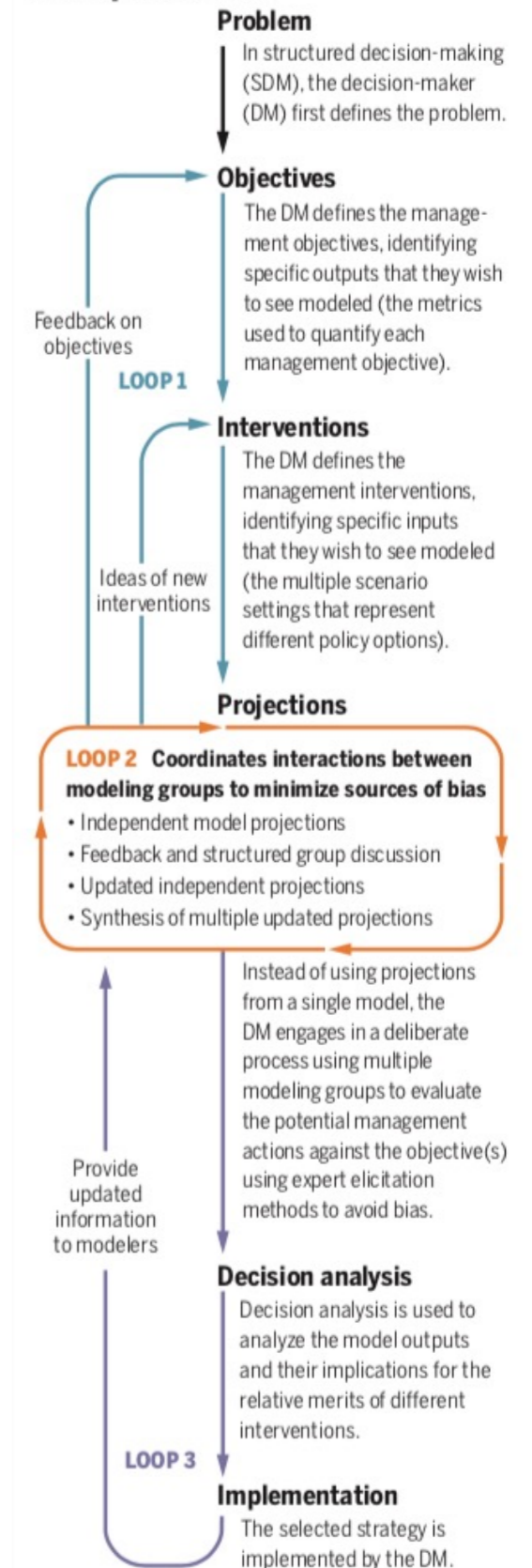
Combining Deliberative and Staticized Groups

“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>

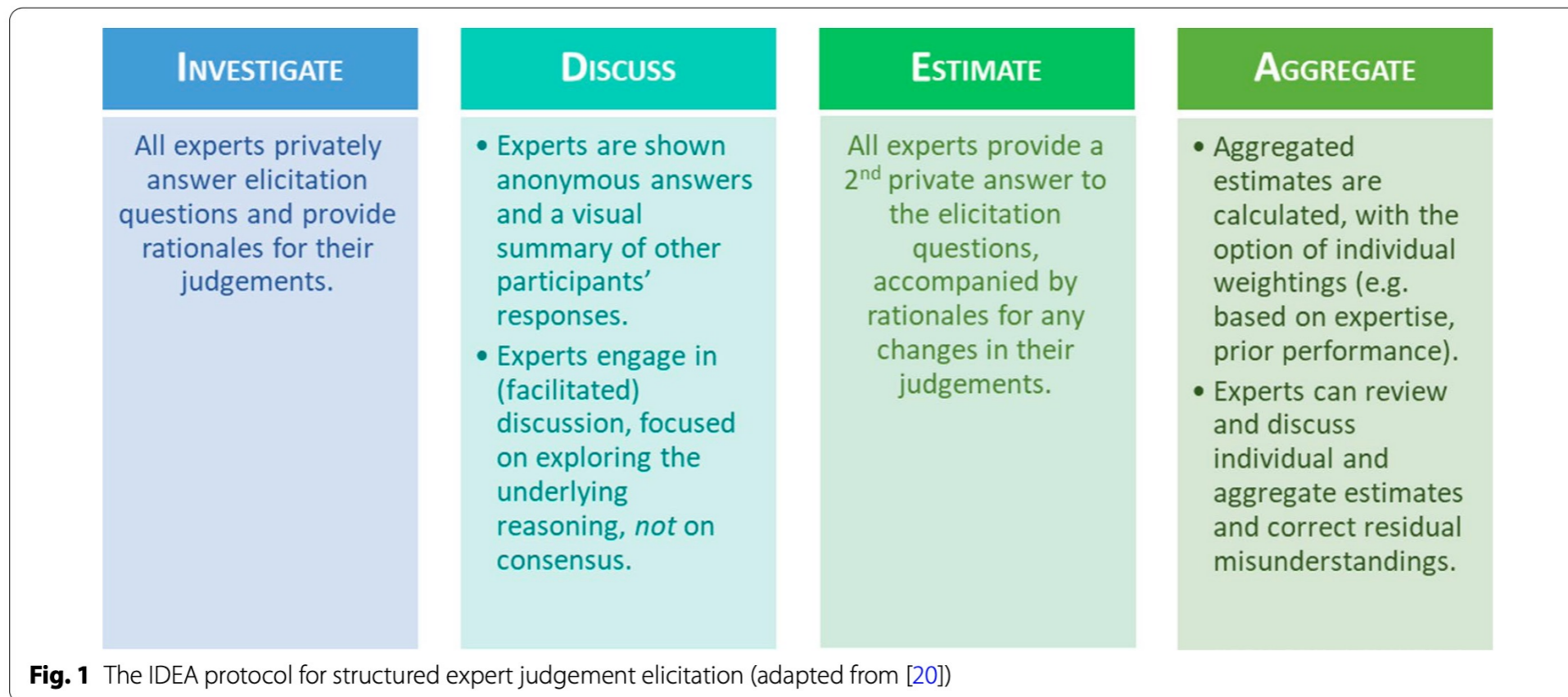
Making the most of multiple models



Exercise: 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.



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>

Have a good week and see you next Monday!
