American Psychological Association

American Psychologist

© 2019 American Psychological Association
0003-066X/19/$12.00

2019, Vol. 74, No. 3, 290–300
http://dx.doi.org/10.1037/amp0000429

From Discipline-Centered Rivalries to Solution-Centered Science: Producing Better Probability Estimates for Policy Makers

Barbara A. Mellers and Philip E. Tetlock
University of Pennsylvania

From 2011 to 2015, the U.S. intelligence community sponsored a series of forecasting tournaments that challenged university-based researchers to invent measurably better methods of forecasting political events. Our group, the Good Judgment Project, won these tournaments by balancing the collaboration and competition of members across disciplines. At the outset, psychologists were ahead of economists in identifying individual differences in forecasting skill and developing methods of debiasing forecasts, whereas economists were ahead of psychologists in designing simple market mechanisms that distilled predictive signals from noisy individual-level data. Working closely with statisticians, psychologists eventually beat the markets by producing better probability estimates that funneled top forecasters into elite teams and aggregated their judgments using a log-odds formula tuned to the diversity of the forecasters. Our research group performed best when team members strove to get as much as possible from their home disciplines, but acknowledged their limitations and welcomed help from outsiders.

Keywords: prediction, forecasting, tournaments, prediction markets, prediction polls

Prior to the path-breaking work of Kahneman and Tversky (Kahneman & Tversky, 1979; Tversky & Kahneman, 1974), few economists paid much attention to psychology. If they thought about it, they might concede that the two core primitives in microeconomic theories of decision-making—probability and utility—bear a family resemblance to beliefs and attitudes in the social psychological literature. But most saw little reason to dwell on the psychological underpinnings of economics, less still neuropsychological ones. Much has changed. Prominent economists, including current and likely future Nobelists, have explored psychological literatures in depth (Akerlof & Schiller, 2010; Fehr & Rangel, 2011; Laibson, 2001; Thaler & Sunstein, 2009; Camerer, 2003). A casual survey of the economic literature reveals thousands of references to psychological constructs, especially, but not exclusively, focused on heuristics and biases, prospect theory, emotions, and fairness (DellaVigna, 2009; Rabin, 1998).

Psychologists who work with a traditional reductionist model of science might see this trend as a belated movement toward the truth. In this view, there is a good reason why knowledge transfers between psychology and economics have been largely one directional. Psychology is the science of human behavior. Economics is the science of allocating scarce resources among human beings enmeshed in complex webs of interdependence. It is natural that the knowledge transfers flow predominantly from the science at the micro, individualist level of analysis to the science at the macro, societal level of analysis.

A casual survey of the psychological literature reinforces this view. Economic constructs widely discussed by psychologists are vastly out-numbered by psychological constructs widely discussed by economists. The conspicuous exception is the field of judgment and decision making which has borrowed normative standards from microeconomics, such as expected utility theory, to organize comparisons between how people should make decisions and how they do make decisions (Baron, 2008).

Our aims in this article are fourfold. First, we make an obvious but often hard-to-accept point. Interdisciplinary collaborations work best when each side starts from the

Editor's note. This article is part of a special issue, "Multidisciplinary Research Teams: Psychologists Helping to Solve Real-World Problems," published in the April 2019 issue of *American Psychologist.* Robert W. Proctor and, Kim-Phuong L. Vu served as guest editors, with Elizabeth A. Klonoff as advisory editor.

Authors' note. Barbara A. Mellers, Department of Psychology and Department of Marketing, Wharton School of Business, University of Pennsylvania; Philip E. Tetlock, Department of Psychology and Department of Management, Wharton School of Business, University of Pennsylvania.

Correspondence concerning this article should be addressed to Barbara A. Mellers, Department of Psychology, Solomon Labs, University of Pennsylvania, 3720 Walnut Street, Philadelphia, PA 19104. E-mail: mellers@wharton.upenn.edu

Barbara A. Mellers

assumption that knowledge transfers should be a two-way street. We stress "hard to accept" because it is not yet fashionable in our field to concede that psychologists have much to learn from economists. We find that a good way to overcome the dismissive discipline-grounded stereotypes that scientists have been known to form of each other is to incentivize investigators to join forces in solving a high-stakes problem that no single field can solve on its own. This prescription is in the spirit of Allport's (1954) recipe for reducing intergroup conflict—as well as in the spirit of what the Intelligence Advanced Research Projects Activity (IARPA), the research and development branch of the Office of the Director of National Intelligence, did in 2011 when it launched a 4-year series of geopolitical forecasting tournaments. The challenge to the funded research groups was to assemble whatever interdisciplinary coalition they thought could do the best job of extracting accurate geopolitical forecasts from a large, dispersed crowd of volunteers. IARPA did not care whether economists or psychologists or computer scientists or statisticians "won." The horse race was the means to an end: the invention of better methods of estimating the probabilities of events of national security interest.

Second, we present some empirical fruits from our group's efforts to test the merits of psychological versus microeconomic approaches to improving human forecasting as well as to develop viable hybrids. The economists' favored approach, the prediction market in which people bet on outcomes, had an initial edge over the psychologists' preferred method, prediction polls, in which participants make probability judgments. But this edge only held when we used a crude method of aggregating probability judgments—simple averaging. That edge disappeared when we used more sophisticated aggregation algorithms, such as a weighted-averaging strategy that favored the most recent forecasts of the best forecasters and then fine-tuned the aggregate as a function of the diversity of the crowd.

Third, we take up the challenges of managing an interdisciplinary team in which we cannot safely assume mutual respect holds sway. Psychologists and economists do indeed harbor less-than-flattering stereotypes of each other—and achieving the right mix of competition and collaboration in our group was nontrivial. The analogy to sports teams is apropos because we were competing in a tournament and had the shared objective of winning, a superordinate goal repeatedly shown to reduce stereotypes (Pettigrew, 1998). But intragroup rivalries are hard to avoid and, as principal investigators, we had to steer intragroup competition in productive directions.

Fourth, we discuss why IARPA invested so much (tens of millions of dollars) in probability-estimation tournaments. In many fields, including intelligence analysis, there is sharp resistance to quantifying beliefs. Professionals often prefer to state their hunches in vague-verbiage form: "distinct possibility" or "reasonable chance." Such phrases reduce the risk of "looking wrong" because they are open to so wide a range of after-the-fact interpretations (Wallsten, Budescu, & Zwick, 1993). But political safety carries a steep price tag. Vague-verbiage forecasting prevents us from getting the precision-accuracy feedback essential for learning to make well-calibrated judgments. How could we discover whether events we saw as 70% likely actually occurred 70% of the time if we cannot reliably reconstruct what we meant by "distinct possibility"? The next interdisciplinary challenge should focus on creating "psychological safe" organizations (Edmondson & Lei, 2014) in which people feel freer to acknowledge their fallibility and accept the fact that mistakes are inevitable in an uncertain world.

Structuring Interdisciplinary Collaboration to Improve Foresight

We define predictions or forecasts as subjective beliefs about unknown outcomes—and see improving forecasting as central to better decision making in all spheres of public policy and indeed life. To make a good decision, normative economic theories tell us to evaluate the alternatives by assessing the utilities of the possible future outcomes and weighting them by the likelihoods of their occurrence (Baron, 2008). Not surprisingly, the quest to boost accuracy is a recurring theme in virtually all professions. Meteorologists try to forecast the weather; central bankers, macroeconomic trends; intelligence analysts, threats to national security. Errors can have tragic consequences. For instance, would the U.S. Senate have authorized military force against Iraq if moderate Senators in both parties had be-

Philip E. Tetlock

lieved the probability of weapons of mass destruction to be lower than the extremely high-confidence probabilities in the national intelligence estimates? Counterfactuals of this sort resist precise answers (Art & Jervis, 2016; Tetlock & Gardner, 2015), but it would be surprising if downsizing from a 99% probability to an 80% or 70% or 60% did not undercut support to some degree.[1] And it is reasonable to posit that, in the long run, more accurate probability estimates will help democratically elected officials choose policies more in alignment with their values. One does not need to improve the accuracy of probability judgments informing multitrillion-dollar decisions by much to justify multimillion-dollar research investments.

Between 2011 and 2015, IARPA sponsored four geopolitical forecasting tournaments to improve human predictions of high-stakes outcomes. Five university research groups competed to invent the best methods of eliciting and aggregating human forecasts where "best" was defined as most accurate. Two features of IARPA's approach merit emphasis.

First, although IARPA was agnostic about "who won," it was particularly curious as to how economic approaches would fare because the intelligence community had already "bet on" prediction markets by establishing their own internal market for use by professional analysts who had security clearances and access to top secret information (Goldstein, Hartman, Comstock, & Baumgarten, 2018).

Second, IARPA did not want to invest a lot in tournaments that could be easily gamed, so it issued a wide range of questions, wide enough to make it impractical for a research group to simply hire specialists who could cover the full bandwidth. The eight examples below—drawn from the 100-plus questions posed in each of the 4 years—illustrate how far-ranging the questions were as follows:

> "Will any country officially announce its intention to withdraw from the Eurozone before April 1, 2013?"
>
> "Will the number of registered Syrian conflict refugees reported by the United Nations High Commissioner for Refugees exceed 250,000 at any point before April 1, 2013?"
>
> "Will China seize control of the second Thomas Shoal before January 1, 2014?"
>
> "Will the World Health Organization report any confirmed cases of Ebola in a European Union member state before 1 June 2015?"
>
> "Will the Colombian government and FARC commence official talks before January 1, 2013?"
>
> "On September 15, 2014, will the Arctic Sea ice extent be less than that on September 15, 2013?"
>
> "Will Spanish government generic 10-year bond yields equal or exceed 7% at any point before September 1, 2012?"
>
> "Will either the French or Swiss inquiries find elevated levels of polonium in the remains of Yasser Arafat's body?"

We were the principal investigators of the group that devised the most accurate sets of forecasting methods each year. Our group consisted of computer scientists, statisticians, behavioral economists, and psychologists. Computer scientists developed the web platform used to launch forecasting questions, elicit predictions, score the forecasters, and provide them with feedback. Psychometricians and statisticians created aggregation algorithms for combining forecasts. Behavioral economists designed the prediction markets, monitored trading behavior, and analyzed pricing signals. Psychologists selected cognitive-ability and cognitive-style measures to predict individual differences in forecasting accuracy, explored the effects of social interaction among teams, developed training modules to instruct forecasters in probabilistic reasoning, and instituted "tracking" to set better forecasters apart.

[1] We recognize that policy makers will sometimes be extremely probability insensitive. Vice President Cheney is reported to have said that, even if the probability of weapons of mass destruction had been as low as 1%, he would have supported the invasion of Iraq because he was now looking at the potential threat through a new lens—that of 9/11 (Suskind, 2006). A great advantage of forecasting tournaments is that they press all of us to be more candid about distinguishing our fact and value judgments in policy debates.

From the outset, there was cross-disciplinary agreement on a key point: Composite forecasts derived by aggregating diverse views among forecasters will usually be more accurate than the forecasts of the designated subject-matter expert (Armstrong, 2001; Tetlock, 2005). The power of this idea, which Surowiecki (2005) popularized in *The Wisdom of Crowds* has been demonstrated in numerous scientific papers (Budescu & Chen, 2014; Mannes, Soll, & Larrick, 2014; Ray, 2006). Averaged forecasts outperform the majority, often the vast majority, of the individual forecasters from whom the averages were derived—a phenomenon driven by the cancellation of random errors associated with independent judgments.

But there was deep-rooted disagreement over how best to elicit and aggregate judgments. Economists tended to organize their thinking around one big, elegant idea. The fastest way to discover the truth is to rely on self-correcting market mechanisms in which people buy or sell shares of futures contracts that pay a fixed amount if an event occurs and nothing otherwise. The price at which a contract trades (the market equilibrium) is, in effect, a collective probability estimate that the event will happen.[2]

To illustrate, imagine a contract that paid $100 if Donald Trump were to be reelected as president in 2020 and $0 otherwise. If that contract were trading at $40, we would say that, given the available information set, the market predicts that Trump has a 40% chance of reelection. Someone who thinks the probability is higher should buy shares, which will increase the price in accord with the law of supply and demand. If someone else thinks the probability is lower, that person should sell shares, which will decrease the price. The economic approach is reminiscent of the parsimony of mid-20th-century behaviorism in psychology, an austere input–output model, with no need to explore messy details about mediating mechanisms, such as how traders in markets balance conflicting arguments about an event or avoid market biases (Kendler, 1987).

By contrast, the psychologists were interested in those messy details. They adapted psychometric models to spot better forecasters (e.g., item response theory; Bo, Budescu, Lewis, Tetlock, & Mellers, 2017; Merkle, Steyvers, Mellers, & Tetlock, 2015, 2017), conferred special recognition on the best forecasters (dubbed *superforecasters*), concentrated them in elite teams (Mellers et al., 2015), designed cognitive-debiasing exercises that summarized the insights of the best forecasters, and conveyed those lessons to all forecasters (Chang, Chen, Mellers, & Tetlock, 2016). The statisticians developed algorithms to aggregate forecasts, including the combination weighted model (Budescu & Chen, 2014; Chen, Budescu, Lakshmikanth, Mellers, & Tetlock, 2016), a minimum distance robust regression model (Cross, Scott, Ramos, Mellers, & Tetlock, 2018), and a log-odds extremizing rule (Baron, Unger, Mellers, & Tetlock, 2014; Satopaa, Baron, et al., 2014, Satopaa, Jensen, et al., 2014).

From a narrowly microeconomic perspective, there is little to be gained from these efforts. Why bother using complex psychometric instruments to capture individual differences in forecasting skill when the market will automatically identify those with skill as the richest traders? Why bother designing opinion-aggregation algorithms that up-weight better forecasters when the market ensures that richer traders will be the bigger market movers? And why bother developing methods of teaching forecasters about cognitive biases and basic concepts of probability when the market will use bankruptcy to weed out the innumerate? In this worldview, psychologists complicate things unnecessarily. Prediction-market researchers can sit back and watch traders do all the heavy cognitive lifting, all the relentless second-guessing of each other's buy-and-sell decisions that directs the market ever closer to its equilibrium price for the day.

The most influential formal argument for market aggregation is the efficient markets hypothesis (Fama, 1970). Prediction markets boost accuracy because they incentivize traders to seek out information and use it to guide their buying and selling whenever they think the market price is wrong. If someone knows something that others do not, that person can and should profit from it. Traders are competitive, and competition stimulates effort and engagement ("skin in the game").[3] The market is an opinion-aggregation system that seamlessly conducts exchanges and generates a dynamic flow of pricing signals that represent traders' best collective guesses of outcome probabilities.

The efficient markets hypothesis asserts that, for reasonably deep, liquid markets, the market price should reflect all information available to traders at any given moment. Setting aside insider trading, no one should be able to consistently out-predict the market. And it is naïve even to ask which method works better—prediction polls or markets. Prediction markets must be the optimal aggregator. If the market loses to a polling technique, it is merely a sign that the market was poorly designed or undercapitalized. Empirical results can shed light on how inept the researchers were, but not on the underlying logical truth that deep, liquid, markets are the most efficient engine for measuring and reducing uncertainty.

Prediction markets do indeed have an impressive track record on a wide array of outcomes, from U.S. elections (Iowa Electronic Markets) to box office revenues for newly released films (Hollywood Stock Exchange, PredictIt, and

[2] However, research suggests that markets with hypothetical cash are also quite accurate (see Servan-Schreiber, Wolfers, Pennock, & Galebach, 2004).

[3] We cannot rerun the 2016 election to determine whether Trump's victory actually was unlikely—and we just happen to inhabit an unlikely world.

Predictwise). Firms from Google to Boeing have used prediction markets that challenge employees to second-guess each other's guesses about meeting company deadlines, reaching levels of product quality, and predicting consumer demand (e.g., Cowgill & Zitzewitz, 2015). Prediction markets have outperformed experts on company sales projections (Plott & Chen, 2002), entertainment-beat journalists on Oscar winners (Pennock, Lawrence, Nielsen, & Giles, 2001), and professional macroeconomists on macroeconomic indicators (Gürkaynak & Wolfers, 2006).

Prior to the IARPA tournaments, however, the literature shed less light on a key question: Do prediction markets beat polls, all else equal? There were some signs this might be true (e.g., Leigh & Wolfers, 2006), but the bulk of the evidence suggested that differences were small and unreliable (e.g., Graefe & Armstrong, 2011; Rieg & Schoder, 2010). Valid tests between methods must be conducted over long periods of time due to the enormous volatility in the real world. Consider that, at midnight on November 7, 2016, the evening before the U.S. presidential election, Nate Silver's poll of polls predicted a 71% chance of a Clinton victory. A major prediction market—the Iowa Electronic Markets—assigned a 69% chance of a Clinton win, and Predictwise went as high as 85%. In the eyes of the public, neither method seemed to work well, although neither approach was technically "wrong" because a probability, p, of an event implies a nonzero probability, $1 - p$, that the event will not occur.

Our research group was divided, along expected disciplinary lines, over which method would yield better forecasts. We decided to run a series of experiments pitting prediction markets against prediction polls. We did agree that, regardless of the experimental condition to which people were randomly assigned, a fair test of the maximum potential of each method would require that all participants get prompt, clear performance feedback essential for learning (Erev & Roth, 1998; Hogarth, 1981). This feedback took the form of earnings in prediction markets and Brier scores in prediction polls.

Off to the Races

IARPA's forecasting tournaments were an unprecedented opportunity to compare the two most promising methods in repeated, large-scale randomized controlled trials. Forecasting questions were launched throughout the year and remained open as little as 1 week or as long as 2 years, with an average of 3 months. Prediction markets are well designed for such questions. Traders can enter and exit the market as often as they wish to buy or sell shares. But with polling methods, questions are typically asked once and they are often about preferences, not predictions. We needed a dynamic polling system that participants could enter and exit to update their probability predictions as easily as traders could in the market. The unsung heroes of our group, the programmers, made this possible.

The futures contracts in the prediction markets ranged in value from \$0 to \$1. If the event occurred, shares became worth \$1, otherwise, \$0. Participants started the tournament with \$100 in hypothetical cash and were given small-cash infusions throughout the tournament to enhance market liquidity when trading fell below a threshold volume (a procedure akin to what the Federal Reserve does to expand money supply to stimulate economic activity during slowdowns).

Although prediction polls quantify uncertainty using probability judgments and markets quantify uncertainty using prices, the prices can be readily converted into probability estimates which let us compare the two methods on a widely used accuracy metric known as the Brier score (Brier, 1950) which, for a given question, was computed as follows:

$$S = 1/(j)\sum_j \sum_i (f_{ij} - o_{ij})^2,$$

where S is the Brier score, f_{ij} was a forecast with i outcomes on day j, and o_{ij} is the outcome (scored as 0 if the event did not occur or 1 if it did). We then averaged Brier scores over questions. Scores ranged from 0 (*best*) to 2 (*worst*). Brier scoring punishes overconfidence quite severely. For example, if a forecaster said a binary event was 90% likely and the event did not occur, the Brier score would be 1.62. Furthermore, Brier scoring rewards justified decisiveness generously. If a forecaster said the event was 90% likely and it did occur, the Brier score would be a close-to-omniscient .02. The Brier scoring rule is "proper" in the sense that forecasters get the best score if they report their beliefs truthfully (Gneiting & Raftery, 2007)—and do not distort their beliefs to serve nonepistemic functions, like wishful thinking or defensive pessimism. Brier scoring treats under- and overestimation of events in exactly the same way, as equally erroneous.

A key member of our group, Pavel Atanasov, took the lead on two consecutive, 9-month experiments comparing prediction markets to polls, using 3,500 college-educated volunteers (Atanasov et al., 2017).[4] The markets were continuous double auctions in which people traded by placing buying prices and selling prices into an order book, which displayed the six highest buying prices and six lowest selling prices. Price history was public information.

Extending this work, another team member, Jason Dana, launched an even more demanding study in which people participated in a prediction market and a prediction poll simultaneously (Dana, Atanasov, Tetlock, & Mellers, in press). In this repeated-measures design, participants logged onto a prediction market. But before they could trade, they

[4] There were other conditions in the experiments that we do not discuss. See Atanasov et al. (2017) for details.

had to judge the probability of the event. This design ensured that participants could see the same information while making exchanges and assigning probabilities, the last trading price and the history of trades in the order book.

In all three experiments, prediction markets beat prediction polls when forecasts were combined using a simple mean. One could call this a "victory" for the economists. But we also compared prediction markets to prediction polls with a more sophisticated weighted averaging algorithm that gave more weight to the most recent forecasts of the most accurate forecasters and that then recalibrated the weighted-average forecast, p, as follows:

$$p^* = p^a/[p^a + (1 - p)^a],$$

where p^* is the extremized aggregate probability estimate, and a, the exponent, is the recalibration parameter. When $a = 1$, there is no recalibration. When $a > 1$, transformed values, p^*, become more extreme and are pushed closer to 1 if $p > .5$ or closer to 0 if $p < .5$. We performed this transformation because averages typically do not reflect all of the information in the individual estimates, even when individuals are well calibrated (see Baron et al., 2014). To estimate the value of a, we fit the model to forecasts from a previous year. Using this aggregation rule, team-prediction polls beat the prediction markets in two experiments (Atanasov et al., 2017), and tied in the third (Dana et al., in press). If there is an edge, it favors the prediction polls.

How did this happen? There is an intuitive way to think about the psycho-logic underlying the winning algorithm. Imagine a forecasting tournament organized around the 2020 presidential election. Different schools of thought drawing on different methods and conceptual frameworks compete to generate the most accurate predictions. These include advocates of prediction markets, polling methods, meta-analytic methods of aggregating polls and polimetric models of elections that focus on systemic variables like the economic growth and incumbency status (Abramowitz, 1988; Fair, 1978; Murphy, 1988).

Suppose each of these diverse methods independently converges on a 35% to 45% range of probabilities of a Trump victory in 2020. Our algorithm would tell us to make that probability even more extreme, pushing it lower, say, 20% to 30%. Conversely, suppose the methods converged on a 55% to 65% range. The algorithm would now tell us to "extremize," pushing it even higher, up to 70% or 80%. The core idea is that when we see forecasting methods whose judgments are usually uncorrelated suddenly agree, the algorithm tells us to treat that agreement as especially informative. This idea was demonstrated by Wallsten, Budescu, Erev, and Diederich (1997) and later proven under reasonable assumptions by Wallsten and Diederich (2001).

The accuracy of prediction polls was also boosted from three behavioral interventions: training, teaming, and tracking (Mellers et al., 2014, 2015). Our training manipulation was minimalist, a 45-min tutorial that covered the basics of probabilistic reasoning and cognitive biases, encouraged the use of base rates and reference classes, and provided fast-and-frugal heuristics, such as averaging when there was more than one plausible base rate. In a succession of randomized controlled studies, Chang et al. (2016) showed that this tutorial boosted performance in each of the four years by something between 6 and 11%.

Our teaming manipulation was also simple. We randomly assigned forecasters to work alone or in teams of 10 to 15 members. The argument for working alone is statistical; errors should cancel out when we combine independent forecasts. The argument for working together is psychological; teams should benefit from sharing information and increasing engagement but only if they can avoid the twin perils of factionalism and groupthink. We told teams about this balancing act and the teams managed to pull it off. They quite consistently surpassed independent forecasters (Mellers et al., 2014).

Finally, tracking involved culling out the best performers (the top 2% at the end of each year), calling them superforecasters and assigning them to elite teams to work together the following year. This proved the most potent of all the behavioral manipulations. Superforecaster teams were between 29% and 38% more accurate than other teams and also outperformed prediction markets.

The behavioral economists in our group were understandably curious whether the behavioral interventions might also boost the accuracy of prediction markets. For two years, we created, tested and refined training modules for prediction markets that offered advice on how to adjust trading positions and tips for avoiding biases. We also ran a special prediction-market condition in which traders were incentivized to collaborate within teams and compete across teams. And we tracked better traders from the previous year, called them "supertraders," and put them into an elite "supermarket." The yield from these interventions was however meager, with no effects reaching statistical significance. What worked in the polling world did not transfer to the market world.

A cross-disciplinary team did however eventually discover a way to boost market performance by combining self-report and market-pricing data (Dana et al., in press). The working hypothesis was that it should be possible to tamp down distortions in prediction markets, such as the favorite-longshot bias (Manski, 2006) in which traders overvalue longshots and undervalue favorites, by taking into account human judgments that, for various reasons, do not get translated into trades.

To test this idea, we compared the accuracy of prediction markets and polls on short versus long duration questions and found that for longer questions, prediction polls were more accurate. Probability judgments were more accurate

than prices when the resolution of the question was months away, but prices were more accurate than probabilities right before the question resolved (Atanasov et al., 2017; Dana et al., in press). Traders apparently disliked tying up their assets on questions of longer duration which created an opportunity for polls to complement markets.

We also found that polls complement markets when trading volume is thin (Dana et al., in press), and there is a wide gap between the lowest offered selling price and highest offered buying price. If the price is too high and the bid too low, then trading cannot occur. When volume is low, the market price will be less informative and presumably less accurate. In brief, we began identifying the conditions under which we could boost the predictive power of the market by "hybridizing"—by augmenting price signals with self-report judgments.

A final wrinkle deserves note, and we believe it is the most important result from a public-policy perspective. U.S. government researchers compared our best performing prediction polls and markets in open-to-the-public tournaments to an internal prediction market that was operated behind a veil of security clearances by the intelligence community (Goldstein et al., 2018). This market is known as the Intelligence Community Prediction Market. Drawing on 130 questions that were also posed to our forecasters, Goldstein et al. found that professional intelligence analysts making trades on the Intelligence Community Prediction Market were (a) significantly more accurate than the unweighted average of our prediction poll; (b) roughly as accurate as our prediction market; and (c) less accurate than our superforecasters and prediction polls that were aggregated using our best performing log-odds, extremizing algorithm.

It understates the importance of these results to say that Goldstein et al. (2018) replicated the underperformance of the prediction market in a high-stakes institutional setting. The Goldstein et al. results raise serious questions about how to improve the probability estimates flowing into public policy. Were professional intelligence analysts less accurate than college-educated amateurs because (a) access to classified information sometimes hurts, rather than helps, (b) forecasting accuracy in the market was not highly valued inside the intelligence community, and (c) the market was not functioning optimally because wealth was too concentrated among a small number of traders? We do not know. But we do know that the probabilities intelligence analysts assign to outcomes matter, and the answers to these questions are of more than academic interest.

Sustaining an Interdisciplinary Team and Overcoming Disciplinary Stereotyping

The obvious analogy for our research group was that of a "sports team" whose members play complementary roles, and whose success depends on fostering both cooperation within the group and competition with out-groups. After all, IARPA had explicitly primed that analogy. It framed the entire project as a "tournament," from which a winner would emerge and eventually, the other teams would be "down-selected." So, it was natural to see our job as principal investigators as a blend of the general manager, charged with managing budgets, setting high-level goals, recruiting essential talent, and finding the resources needed to keep the operation afloat, and of the coach, charged with allocating talent to tasks in response to the emergencies of the day, facilitating interactions among players, building morale, and dealing with the inevitable tensions that arise when one part of the team feels it has gotten the short end of the tradeoff stick.

Managing these challenges is hard in interdisciplinary research because professionals from different disciplines have different conceptions of ideal work products, as well as different publication norms. For instance, pretenure psychologists felt greater pressure to publish more quickly than did pretenure economists and statisticians, who were more inclined to wait and see what patterns emerged across years than to publish data from a single tournament year. In the wake of the replication crisis, many psychologists are now coming around to the economists' view that excessive publication expectations may have incentivized questionable research practices.

Coordination is also hard when colleagues from different disciplines harbor unflattering stereotypes of each other. To put it bluntly, economists tended to doubt the depth of psychologists' understanding of statistics (econometrics courses are more demanding than most psychology graduate statistics classes). Economists also suspected psychologists harbored an antimarket bias that blinded them to the potential power of prediction markets. Conversely, psychologists tended to see economists as overconfident, insensitive to the importance of individual differences, and prone to view markets as a one-size-fits-all solution to the problem of better informing public policy.

A little knowledge of the literature on intergroup conflict proved helpful here. Tensions rise when people see themselves in competition for a static or shrinking pie (Prediger, Vollan, & Herrmann, 2014). We confess to using this insight in a fashion that our statistician colleagues saw as a tad opportunistic. At the beginning of the first year, when only a handful of forecasting questions had resolved, IARPA released very preliminary standings that showed the Good Judgment Project taking a lead that looked large, but was based on far too small a sample size to be given much statistical weight.

We faced a choice. We could have heeded our statistician superegos and cautiously downplayed the results as "inconclusive" or we could heeded the classic coaching advice of building up morale by playing up the results as a "first win" and a harbinger of bigger victories to come (Katz, 2001) or

we could have adopted the middle course. We took the compromise. We sent out an optimistic message to the forecasters that, although the data were still fragmentary, their hard work was in our opinion already starting to pay off—and a message to our programmers that their efforts to ensure that forecasts were aggregated and transmitted in a glitch-free manner were hugely appreciated. The challenge now became building on this early momentum, which did indeed happen as the next several waves of data transformed a weak trend into a strong one.

We also tried to overcome interdisciplinary tensions through another classic social psychological tactic. We encouraged lots of formal and informal communication among team members, communication that was usually concentrated within a specialized function (e.g., improving the user-friendliness of the website for collecting data in forecasting polls and prediction markets or boosting the computational reliability of our aggregation algorithms, or fine-tuning forecaster training), but that sometimes spread across multiple functions (e.g., developing better methods for helping forecasting teams share information or giving individuals and teams faster leaderboard feedback).

Formal communications occurred in two forums: via conference calls, which tended to focus on specific problems in need of quick action, and via project seminars, which tended to explore bigger-picture questions. Each forum required its own ground rules. The conference calls required great self-control. The rules were that (a) people would remain silent unless they felt they had important suggestions that, if not expressed, would cause the conversation to go onto the wrong track; (b) when people did talk, they should strive to be succinct; and (c) calls should always end at a designated time (signaling that time was a high-value commodity that should not be frittered away). But even our strict ground rules allowed for exceptions. For instance, we encouraged call organizers to ask more introverted team members to speak up more than they might naturally have done, a wise decision in light of the correlation in our team between shyness and possession of crucial information on the operation of essential data-analytic systems (Cain, 2013).

The highest profile meetings were the in-person project seminars at the University of Pennsylvania, with call-in capacity for collaborators at the University of California, Berkeley and elsewhere. We structured many of these seminars along the lines of "adversarial collaborations," a method of scientific dispute resolution developed by Kahneman (2011) and his collaborators. We scheduled these seminars to occur twice a month, with exceptions for emergencies and visiting scholars or government officials.

These exercises served two functions. The first was a managerial one in which the goal was to channel intragroup rivalries down constructive rather than destructive paths. Following Kahneman's rules for adversarial collaboration (Mellers, Hertwig, & Kahneman, 2001), one first needs to demonstrate that one understands the perspective of the other side and can summarize its arguments so well that the other side feels fully and fairly represented. This sort of perspective-taking exercise tends to reduce caricaturing and stereotyping—and build trust (Weyant, 2007). It also lays the basis for the next phase of the conversation, the joint design of studies to determine whose position is closer to the empirical truth. In this phase, we learn who is willing to put professional–reputational capital on an idea and carry out the difficult task of testing it.

The second function of adversarial collaborations was to improve the accuracy of our mainstay methods, polls and markets. It is hard to trace exactly the origins of ideas and to identify precisely whom to credit for what. But adversarial collaborations played a major role in developing virtually all of the tools and interventions described earlier for winning the IARPA tournaments. For instance, psychologists were responsible for improvements in the prediction markets (especially the experiment that boosted performance by measuring self-report probabilities as well as trading prices), and economists were responsible for improvements in probabilistic-reasoning tutorials (exercises that helped prediction poll participants).

To summarize, our group-dynamics recipe for interdisciplinary collaboration includes (a) a superordinate goal that all team members endorse (a tried-and-true method of checking stereotyping; Gaertner et al., 2000) and (b) a within-team accountability mechanism—in our case, adversarial collaboration—that incentivized perspective-taking and good-faith negotiations with intellectual rivals in the service of advancing science. Beyond that, we would add two more generic managerial prescriptions, each with massive research support: (c) a commitment to procedural justice, which required the principal investigators to acknowledge our potential biases as psychologists and to give a full hearing to dissenting voices, especially when the dissent was coming from nonpsychologists, and (d) a commitment to setting and meeting stretch goals, which had to be continuously updated for two reasons. First, the tournament competition created constant pressure to keep bringing down Brier scores (improving accuracy). Second, the tournament imposed a daily deadline. Our group had to submit our official probabilistic forecasts before 9 a.m. EST, each day for 4 years, the sort of production schedule one sees in financial institutions but rarely in academia. It did not matter how clever the research strategies for generating probability estimates were if we could not count on accurate, timely computations and transmissions of those estimates. It may seem pedestrian, but it was the undoing of at least one of our research competitors in the IARPA tournament. We are keenly aware that our programmers were essential for making all the discoveries from this project.

What Next?

One of the pleasures of collaborating with colleagues in other disciplines is that you learn to see your own discipline's maintained assumptions from the perspective of outsiders. To get accurate forecasts, economists preferred market mechanisms, whereas psychologists preferred self-report measures of beliefs. These preferences led us down very different research paths. Psychologists wanted on prediction polls with probability judgments, whereas economists wanted on market mechanisms. They saw less need to design supplementary mechanisms to identify top performers and sort them into teams, or to check groupthink inside teams, or to train average performers to adopt the lessons learned by the best performers, or to construct weighted-averaging algorithms for blending the most recent forecasts of the best forecasters and then extremizing in proportion to the diversity of the forecasters aggregated. Market pricing could, in theory, handle all of that, though, in practice, it fell somewhat short. By contrast, psychologists were less enthralled by markets. They saw it as a mistake to use black-box mechanisms when we have ready access to the complex patterns of interaction among thoughtful forecasters working in a data-rich environment.

During the four years of IARPA tournaments, members of our research group had many opportunities to explore each other's blind spots. We gradually got better at improving prediction polls with various behavioral and statistical interventions, but it proved stubbornly hard to improve prediction markets. One possibility is that both methods are approaching diminishing marginal predictive returns. No one ever expected a deterministic world in which Brier scores of zero were possible—and many observers were surprised that Brier scores could be pushed as far down as they were, falling as low as .12 to .14 for the best polling algorithms and .17 to .19 for prediction markets. In this view, we may be reaching the point of irreducible uncertainty—and IARPA should be skeptical of future investments in geopolitical forecasting.

However, we know that any announcements of "mission accomplished" are highly premature. And even if it were true that IARPA had reached the point of diminishing returns on improving geopolitical forecasting, that should not mark the end of collaboration between economists and psychologists. Forecasting only scratches the surface of the much larger project of promoting greater rationality in public policy. The final output of even the best-run prediction market or poll is just a number. Policymakers want more. What is the "story" behind that number? Is there consensus? What are the key drivers of the event? And how can we influence the event? The limitations of purely numerical forecasts should remind us of how much psychologists and economists both ignored in our tournament-treadmill quest to maximize accuracy.

We see improving the reasoning behind forecasting as a domain in which psychologists have a natural advantage. When forecasters are held accountable solely for their skill in out-performing other forecasters—which is what prediction markets do—they have no incentives to share information or help others to sharpen their thinking. Well-designed prediction polls can incentivize both collaboration (e.g., within teams) and competition (across teams).

Economists, however, may have the disciplinary advantage when we leave controlled research settings and enter real-world organizations. A flourishing field inside microeconomics, agency theory, specializes in modeling the ever-evolving cat-and-mouse games between "principals"—those paying for the forecasts—and "agents"—the forecasters. Employers of forecasters in the media, universities, government, or private sector often want "their" forecasters to achieve many goals beyond mere accuracy, including getting attention, being entertaining, playing to the prejudices of key constituencies, and avoiding saying anything that could later prove embarrassing.

Revisit, in this light, the Goldstein et al. (2018) study of professional analysts participating in the intelligence community's official prediction market. One possible explanation for why analysts lost to superforecasters is that analysts have learned to survive in a world that demands juggling multiple goals, whereas superforecasters have learned to survive in an artificial tournament world that requires maximizing one goal, accuracy. For instance, inside the intelligence community, the directionality of prediction errors—under- or overestimating threats—carries consequences. Right after 9/11, underestimating another threat to the U.S. could have been devastating to one's career and indeed the intelligence community. Right after 2003 when it was clear there were no weapons of mass destruction in Iraq, the error of overestimating another threat could have been just as devastating.

In a blame-game world, it is understandable why many professionals oppose the quantification of beliefs (Lanir & Kahneman, 2006; Tetlock & Mellers, 2011). The rational bureaucratic-political response is to retreat into vague-verbiage forecasting –"there is a distinct possibility of a major war"—that makes it virtually impossible to pin analysts down. "Distinct possibility" can take on values as low as 10–20% or as high as 80–90% in readers minds (Tetlock & Gardner, 2015). If the event does occur, analysts can say "I warned you that it was distinctly possible," and if the event does not occur, they can equally rightly say, "I merely said it was possible."

But political safety is cognitively costly. Vague-verbiage forecasting prevents analysts, or indeed any professionals, from getting the feedback they need to become well-calibrated. Our joint work with economist Richard Zeckhauser has shown that there are real returns to precision. It turns out that better forecasters are the ones who make more nuanced distinctions along the probability continuum

(Friedman, Baker, Mellers, Tetlock, & Zeckhauser, 2018) and those distinctions help them more reliably distinguish between 60/40 and 40/60 bets, even 55/45 and 45/55 bets. Granularity in assessments of uncertainty pays off in accuracy.

Practitioners of cost–benefit analysis can readily compute the net value to society of having productions systems—like prediction polls and markets—that generate better probability estimates for policymakers (Sunstein, 2018). But setting up production systems that hold up under real-world pressures will be a huge challenge that will require drawing on many areas of behavioral and social science. It will not be enough just to order analysts to use numbers. Analysts will need to feel that it is psychologically safe to do what "superforecasters" did—and focus, laser-like, on accuracy. That means analysts will need to feel that they can make mistakes of either under- or overestimation of threats and opportunities and resist shading their estimates to please powerful factions that would prefer one answer over the other (Edmondson & Lei, 2014). None of that will be easy. Organizations that speak truth to power are hard to sustain.

References

Abramowitz, A. (1988). An improved model for predicting presidential election outcomes. *Political Science & Politics, 21,* 843–847. http://dx.doi.org/10.1017/S1049096500034235

Akerlof, G., & Schiller, R. (2010). *Animal spirits: How human psychology drives the economy and why it matters for global capitalism*. Princeton, NJ: Princeton University Press.

Allport, G. W. (1954). *The nature of prejudice*. Cambridge, MA: Perseus.

Armstrong, S. (2001). *Principles of forecasting: A handbook for researchers and practitioners*. New York, NY: Kluwer Academic Publishers. http://dx.doi.org/10.1007/978-0-306-47630-3

Art, R., & Jervis, R. (Eds.). (2016). *International politics: Enduring concepts and contemporary issues* (10th ed.). Boston, MA: Pearson.

Atanasov, P., Rescober, P., Stone, E., Swift, S., Servan-Schreiber, E., Tetlock, P., . . . Mellers, B. (2017). Distilling the wisdom of crowds: Prediction markets versus prediction polls. *Management Science, 63,* 691–706. http://dx.doi.org/10.1287/mnsc.2015.2374

Baron, J. (2008). *Thinking and deciding* (4th ed.). New York, NY: Cambridge University Press.

Baron, J., Unger, L., Mellers, B., & Tetlock, P. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis, 11,* 133–145. http://dx.doi.org/10.1287/deca.2014.0293

Bo, E. Y., Budescu, D. V., Lewis, C., Tetlock, P., & Mellers, B. (2017). An IRT forecasting model: Linking proper scoring rules to item response theory. *Journal of Judgment and Decision Making, 12,* 90–103.

Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78,* 1–3.

Budescu, D. V., & Chen, E. (2014). Identifying expertise to extract the wisdom of crowds. *Management Science, 61,* 249–286.

Cain, S. (2013). *Quiet: The power of introverts in a world that can't stop talking*. New York, NY: Penguin.

Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton, NJ: Princeton University Press.

Chang, W., Chen, E., Mellers, B., & Tetlock, P. (2016). Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments. *Journal of Judgment and Decision Making, 11,* 509–526.

Chen, E., Budescu, D., Lakshmikanth, S., Mellers, B., & Tetlock, P. (2016). Validating the contribution-weighted model: Robustness and cost-benefit analyses. *Decision Analysis, 13,* 1–25. http://dx.doi.org/10.1287/deca.2016.0329

Cowgill, B., & Zitzewitz, E. (2015). Corporate prediction markets: Evidence from Google, Ford, and Firm X. *The Review of Economic Studies, 82,* 1309–1341. http://dx.doi.org/10.1093/restud/rdv014

Cross, D., Scott, D., Ramos, J., Mellers, B., & Tetlock, P. (2018). Robust forecast aggregation: Fourier L2E regression. *Journal of Forecasting, 37,* 259–268. http://dx.doi.org/10.1002/for.2489

Dana, J., Atanasov, P., Tetlock, P., & Mellers, B. (in press). Are markets more accurate than polls? The surprising informational value of "just asking,". *Journal of Judgment and Decision Making*.

DellaVigna, S. (2009). Psychology and economics: Evidence from the field. *Journal of Economic Literature, 47,* 315–372. http://dx.doi.org/10.1257/jel.47.2.315

Edmondson, A., & Lei, Z. (2014). Psychological safety: The history, renaissance, and future of an interpersonal conflict. *Annual Review of Organizational Psychology and Organizational Behavior, 1,* 23–43. http://dx.doi.org/10.1146/annurev-orgpsych-031413-091305

Erev, I., & Roth, A. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *The American Economic Review, 88,* 848–881.

Fair, R. (1978). The effect of economic events on votes for the president. *The Review of Economics and Statistics, 60,* 159–173. http://dx.doi.org/10.2307/1924969

Fama, E. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance, 25,* 383–417. http://dx.doi.org/10.2307/2325486

Fehr, E., & Rangel, A. (2011). Neuroeconomic foundations of economic choice—Recent advances. *The Journal of Economic Perspectives, 25,* 3–30. http://dx.doi.org/10.1257/jep.25.4.3

Friedman, J., Baker, J., Mellers, B., Tetlock, P., & Zeckhauser, R. (2018). The value of precision in probability assessment: Evidence from a large-scale geopolitical forecasting tournament. *International Studies Quarterly, 62,* 410–422.

Gaertner, S., Dovidio, J., Banker, D., Houlette, M., Johnson, K., & McGlynn, E. (2000). Reducing intergroup conflict: From superordinate goals to decategorization, recategorization, and mutual differentiation. *Group Dynamics, 4,* 98–114. http://dx.doi.org/10.1037/1089-2699.4.1.98

Gneiting, T., & Raftery, A. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association, 102,* 359–378. http://dx.doi.org/10.1198/016214506000001437

Goldstein, G., Hartman, R., Comstock, E., & Baumgarten, T. (2018). *Assessing the accuracy of geopolitical forecasts from the U.S. intelligence community's prediction market*. Manuscript under review.

Graefe, A., & Armstrong, S. (2011). Comparing face-to-face meetings, nominal groups, Delphi and prediction markets on an estimation task. *International Journal of Forecasting, 27,* 183–195. http://dx.doi.org/10.1016/j.ijforecast.2010.05.004

Gürkaynak, R. S., & Wolfers, J. (2006). *Macroeconomic derivatives: An initial analysis of market-based macro forecasts, uncertainty, and risk* [Working Paper 11929]. Cambridge, MA: National Bureau of Economic Research.

Hogarth, R. (1981). Beyond discrete biases: Functional and dysfunctional aspects of judgmental heuristics. *Psychological Bulletin, 90,* 197–217. http://dx.doi.org/10.1037/0033-2909.90.2.197

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Strauss, and Giroux.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47,* 263–291. http://dx.doi.org/10.2307/1914185

Katz, N. (2001). Sports teams as a model for workplace teams: Lessons and liabilities. *The Academy of Management Executive, 15,* 56–67.

Kendler, H. H. (1987). *Historical foundations of modern psychology*. Philadelphia, PA: Temple University Press.

Laibson, D. (2001). A cue-theory of consumption. *The Quarterly Journal of Economics, 116,* 81–119. http://dx.doi.org/10.1162/003355301556356

Lanir, Z., & Kahneman, D. (2006). An experiment in decision analysis in Israel in 1975. *Studies in Intelligence, 50,* 4.

Leigh, A., & Wolfers, J. (2006). *Competing approaches to forecasting elections: Economic models, opinion polling and prediction markets* [Working Paper 12053]. Cambridge, MA: National Bureau of Economic Research. http://dx.doi.org/10.3386/w12053

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology, 107,* 276–299. http://dx.doi.org/10.1037/a0036677

Manski, C. (2006). Interpreting the predictions of prediction markets. *Economics Letters, 91,* 425–429. http://dx.doi.org/10.1016/j.econlet.2006.01.004

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science, 12,* 269–275. http://dx.doi.org/10.1111/1467-9280.00350

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., . . . Tetlock, P. (2015). Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science, 10,* 267–281. http://dx.doi.org/10.1177/1745691615577794

Mellers, B. A., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., . . . Tetlock, P. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science, 25,* 1106–1115. http://dx.doi.org/10.1177/0956797614524255

Merkle, E., Steyvers, M., Mellers, B., & Tetlock, P. (2015). Item response models of probability judgments: Application to a geopolitical forecasting tournament. *Decision, 3,* 22.

Merkle, E., Steyvers, M., Mellers, B., & Tetlock, P. (2017). A neglected dimension of good forecasting judgment: The questions we choose matter. *International Journal of Forecasting, 33,* 817–832. http://dx.doi.org/10.1016/j.ijforecast.2017.04.002

Murphy, G. (1988). The impact of personal and national economic conditions on the presidential vote: A pooled cross-sectional analysis. *American Journal of Political Science, 32,* 137–154. http://dx.doi.org/10.2307/2111314

Pennock, D., Lawrence, S., Nielsen, F., & Giles, C. (2001). Extracting collective probabilistic forecasts from web games. In *Proceedings of the 7th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 174–183). New York, NY: ACM.

Pettigrew, T. F. (1998). Intergroup contact theory. *Annual Review of Psychology, 49,* 65–85. http://dx.doi.org/10.1146/annurev.psych.49.1.65

Plott, C., & Chen, K. (2002). *Information aggregation mechanisms: Concept, design and implementation for a sales forecasting problem* [Working Paper]. Pasadena: California Institute of Technology.

Prediger, S., Vollan, B., & Herrmann, B. (2014). Resource scarcity and antisocial behavior. *Journal of Public Economics, 119,* 1–9. http://dx.doi.org/10.1016/j.jpubeco.2014.07.007

Rabin, M. (1998). Psychology and economics. *Journal of Economic Literature, 36,* 11–46.

Ray, R. (2006). Prediction markets and the financial "wisdom of crowds." *Journal of Behavioral Finance, 7,* 2–4. http://dx.doi.org/10.1207/s15427579jpfm0701_1

Rieg, R., & Schoder, R. (2010). Forecasting accuracy: Comparing prediction markets and surveys: An experimental study. *Journal of Prediction Markets, 4,* 1–19.

Satopaa, V., Baron, J., Foster, D., Mellers, B., Tetlock, P., & Ungar, L. (2014). Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting, 30,* 344–356. http://dx.doi.org/10.1016/j.ijforecast.2013.09.009

Satopaa, V. A., Jensen, S. T., Mellers, B. A., Tetlock, P., & Ungar, L. (2014). Probability aggregation in time-series: Dynamic hierarchical modeling of sparse expert beliefs. *The Annals of Applied Statistics, 8,* 1256–1280. http://dx.doi.org/10.1214/14-AOAS739

Servan-Schreiber, E., Wolfers, J., Pennock, D., & Galebach, B. (2004). Does money matter? *Electronic Markets, 14,* 1–9. http://dx.doi.org/10.1080/1019678042000245254

Sunstein, C. (2018). *The cost-benefit revolution*. Boston, MA: MIT Press.

Surowiecki, J. (2005). *The wisdom of crowds*. New York, NY: Penguin.

Suskind, R. (2006). *The one percent solution*. New York, NY: Simon & Schuster.

Tetlock, P. (2005). *Expert political judgment: What is it and how do we know?* Princeton, NY: Princeton University Press.

Tetlock, P., & Gardner, D. (2015). *Superforecasting: The art and science of prediction*. New York, NY: Penguin.

Tetlock, P. E., & Mellers, B. A. (2011). Intelligent management of intelligence agencies: Escaping the accountability blame game by signaling commitment to trans-ideological epistemic values. *American Psychologist, 66,* 542–554.

Thaler, R., & Sunstein, C. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. New York, NY: Penguin Books.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131. http://dx.doi.org/10.1126/science.185.4157.1124

Wallsten, T. S., Budescu, D. V., Erev, I., & Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making, 10,* 243–268. http://dx.doi.org/10.1002/(SICI)1099-0771(199709)10:3<243::AID-BDM268>3.0.CO;2-M

Wallsten, T. S., Budescu, D. V., & Zwick, R. (1993). Comparing the calibration and coherence of numerical and verbal probability judgments. *Management Science, 39,* 176. http://dx.doi.org/10.1287/mnsc.39.2.176

Wallsten, T. S., & Diederich, A. (2001). Understanding pooled subjective probability estimates. *Mathematical Social Sciences, 41,* 1–18. http://dx.doi.org/10.1016/S0165-4896(00)00053-6

Weyant, J. (2007). Perspective taking as a means of reducing negative stereotyping of individuals who speak English as a second language. *Journal of Applied Social Psychology, 37,* 703–716. http://dx.doi.org/10.1111/j.1559-1816.2007.00181.x

Received July 15, 2018
Revision received October 17, 2018
Accepted October 21, 2019 ■